AD________________

Award Number: DAMD17-98-1-8044

TITLE: A Computer-Based Decision Support System for Breast Cancer Diagnosis

PRINCIPAL INVESTIGATOR: Zuyi Wang

CONTRACTING ORGANIZATION: The Catholic University of America
Washington, DC 20064

REPORT DATE: September 2001

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
September 2001

3. REPORT TYPE AND DATES COVERED
Annual Summary (1 Sep 00 - 31 Aug 01)

4. TITLE AND SUBTITLE
A Computer-Based Decision Support System for Breast Cancer Diagnosis

5. FUNDING NUMBERS
DAMD17-98-1-8044

6. AUTHOR(S)
Zuyi Wang

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Catholic University of America
Washington, DC 20064

E-Mail: zwang@pluto.ee.cua.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The goal of this project is to develop decision support system for breast cancer diagnosis, treatment option, prognosis, and risk prediction. This project focuses on the development of advanced image pattern analysis in diagnostic imaging and information integration methodology to statistically analyze the distinction between lesion-like normal site and real lesion site. The specific aims of this research project are: (1) image pattern analysis of breast tissue in mammography using both computational features and BI-RADS features provided by radiologist for the prediction of malignancy associated with masses; (2) development of visual presentation methods for radiologists' use in the consultation system; (3) performing a pre-clinical test through an ROC analysis. The clinical goal of this consultation system is to provide scientific tools for doctors to have electronic magnification views, to perform feature analysis of suspected mammographic patterns, to access a large database and investigate clinically similar cases, and to visually inspect the features of a case in various statistical distribution using graphic displays. We have accomplished (1) feature extraction, (2) feature database construction, (3) data mining visual explanation tool development, (4) feature database structure exploration, (5) feature ranking and selection, and (6) classification of mass and non-mass based on selected.

14. SUBJECT TERMS
Breast Cancer Diagnosis, Breast Cancer Patient Database, Decision Support System, Computer-Aided Diagnosis, Visual data exploration, Artificial Intelligence

15. NUMBER OF PAGES
31

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

A Computer-Based Decision Support System for Breast Cancer Diagnosis

Introduction

The goal of this pre-doctoral training project is to develop decision support system for breast cancer diagnosis, treatment option, prognosis, and risk prediction. This system is desired to function as a consultation system for both doctors and patients. This project focuses on the development of advanced image pattern analysis in diagnostic imaging and information integration methodology to statistically analyze the distinction between lesion-like normal site and real lesion site. Based on our intensive observation and experimental evidence, we believe this problem can be better solved through statistical approach. The specific aims of this research project are: (1) image pattern analysis of breast tissue in mammography using both computational features and BI-RADS features provided by radiologist for the prediction of malignancy associated with masses; (2) development of visual presentation methods for radiologists' use in the consultation system; (3) performing a pre-clinical test through an ROC analysis. The clinical goal of this consultation system is to provide scientific tools for doctors to have electronic magnification views, to perform feature analysis of suspected mammographic patterns, to access a large database and investigate clinically similar cases, and to visually inspect the features of a case in various statistical distribution using graphic displays. In the whole period of this research, we have accomplished (1) feature extraction, (2) feature database construction, (3) high dimensional data mining visual explanation tool development, (4) feature database structure exploration using visual exploration tool, (5) feature ranking and selection based on feature database structure exploration, and (6) neural network classifier designing based on selected features through feature database structure analysis.

Overview of Training and Research Accomplishment

1. Research Skill Training and Literature Background Preparation

In the whole period of doctoral training, the development of research skill is very much appreciated through working with the mentors of this, which made it possible for me to continue my research work and further obtain the advanced degree. From the first program for reading and digital mammogram for processing, to the selection of cases, and then to the understanding of fundamental engineering components that are essential for the research, my academic advisor Dr. Yue Wang at The Catholic University of America, my mentors Dr. Shih-Chung Lo and Dr. Matthew Freedmen at Georgetown University Medical Center, provided as much tremendous help as they can. After one year of research work, my insight on research approaches and capability of problem solving have been gradually established and improved. We often discussed and reviewed the primary goal of this project in the research process in order to keep my work in the right direction and give a global view of the all components of CAD. They helped me write better programs for image processing, and discussed the intermediate results of calculation with me for further research planning. Comparing to myself two years ago before working on this project, I see big difference, and I am very grateful.

Under the guidance of Dr. Wang and Dr. Lo, literature and book searching and reading gave me better and broader view of breast cancer and computer-assisted diagnosis (CAD) system research. Through reading engineering textbooks, the fundamental knowledge that is critical to the project is greatly enhanced. The major books I have been reading and using as all-time references are *Neural Network - A Comprehensive foundation* by Simon Haykin, *An Introduction to Signal Detection and Estimation* by H. Vincent Poor, *Elements of Information Theory* by Thomas M. Cover, and,*Statistical analysis of finite mixture distributions* by D. M. Titterington, A. F. M. Smith, and U. E., etc. After searching and technical papers in several major engineering journals, such as IEEE Transactions on Medical Imaging, IEEE Transactions on Neural Network, IEEE Transactions on Pattern Recognition and Machine Intelligence, and Medical Physics, etc., I have collected almost one hundred of relevant papers in order to have an overview of work done by other researchers in this particular area, and also set a start point and direction for my own research. The more I read, the better my capability of understanding and judging other researchers' work. After two years of intensive literature reading, I not only learned many advanced engineering components, but also gradually learned scientific method for problem solving.

2. Research Accomplishments

2.1 Clinical Case and Feature Database Development

2.1.1 Clinical Case Selection

The first step for establishing a feature database was primarily finished in the first and second years, which is case collection and selection that are fundamental and crucial for the further research work. In order to detect suspicious mass regions from a mammogram, we have to be able to find out major differences between mass and non-mass regions so that both mass and non-mass case groups are needed for comparison purpose. The major mammogram sample source that can be accessed and are found proper for the use in this project is ISIS at Georgetown University Medical Center. The ISIS database is constructed by extracting suspicious mass regions from mammograms by licensed radiologists and finally proven by biopsy procedure, from where we obtained 103 cases, among these 71 are mass cases and 32 are non-mass cases. Non-mass cases were purposely selected from normal breast tissue regions with similarity of mass.

2.1.2 Image Feature Extraction

After the preparation of mammogram cases, the next important consideration is to choose features that can be used to distinguish mass and non-mass cases effectively and with high detective rate. The image block was first processed by enhanced segmentation procedure to extract the exact position where a mass may present. The position of the segmented area was then a very useful reference for feature calculation. Many features have been tested by other researchers on their effectiveness for mass and non-mass distinction, and the results have been presented in their most recent papers. Based on literature and medical book searching and reading, primarily we chose nine features, among them are eight texture features and shape feature.

Eight texture features were calculated based spatial gray level dependence matrix, they are energy, correlation, inertia, entropy, inverse difference moment, sum average, sum entropy, and difference entropy. Texture feature, in some scale, may be fairly good for revelation of fine texture differences in images, which cannot be seen by human eyes. They were examined by several research groups for their effectiveness in terms of improvement of CAD performance.

Shape feature, primarily compactness has been used to distinguish non-mass cases from the whole case population in previous study. Through observing hundreds mammograms, shape feature is found to be essential for detecting masses merging in many mass-like normal breast tissues. Most of mass cases are relatively well-defined round objects, however, the overall shape of dense normal breast tissues, such as glandular elements and blood vessels embedded, are often slender rather than round. The simple way of compactness calculation is to divide the area of the segmented area by the square of perimeter of the contour. Therefore the compactness of a perfect circle is one. The closer the object shape is to a circle, the closer the compactness is to one. Compactness calculation is difficult in the first unavoidable and crucial step that is to extract a continuous contour so that a precise perimeter of the segmented area can be calculated. The difficulty of continuous contour extraction comes from the randomness of contour shape and the demand of continuity of contour, even some existing methods proposed for continuous contour extraction in some image processing books cannot cover all possibilities. If only discontinuous contour is needed, the problem becomes very easy since a simple scan of the image can bring us a list of contour pixel coordinates. However, a simple task that can be easily done by human is sometimes very challenging for a computer program. In order to surmount this obstacle, we designed a universal contour extraction method that can deal with all possibilities of position relationship between any pixel and its neighboring pixels, including all kinds of intersections and branches of one contour. The basic strategy of this universal method is that scanning all neighboring pixels of each pixel, memorizing all branches around this pixel, deciding which branch is the right the direction for obtaining a continuous contour, and deleting pixels that have been collected in the contour in order to avoid collecting one same pixel for more than once. Such a method made it possible to precisely calculate the compactness. In the following sections of this report, we will discuss the experimental results that showed that compactness played an important role in the distinction of mass and non-mass cases.

3. Feature Database Structure Exploration and Neural Network Classifier Design

3.1 Visual Data Explanation and Mining Tool Design

Although among many approaches of CAD research, some CAD systems are sophisticated and claimed to have impressive performance, several fundamental issues remain unsolved. For example, Receiver Operating Characteristics (ROC) can provide an overall performance evaluation, but it may not help improve each individual component in CAD system. Furthermore, since machine observer and human observer may not detect the same set of masses, the black box nature of most CAD systems may prevent a natural on-line integration of human and machine intelligence and further upgrade of a CAD system. As a strategic move toward improving CAD design and utility, we developed a visual data exploration and mining tool. Our effort is to (1) provide a visual map of feature database prior to knowledge encoding component so as to evaluate and improve the pre-processing and signature extraction; (2) based on the resulting map to design an optimal classifier best fitted to the particular database structure for knowledge encoding; and (3) combine the map, the classifier output, raw image, and user interface to explore and explain the whole decision making process by both radiologist and CAD system.

3.1.1 Discriminative Projection

Dimension reduction is the first thing on which we spent great effort. There are two major reasons why we have to do dimension reduction: (1) visualization demand (2) cluster separation. Due to the high dimensionality of the feature dataset (in this case, the number of dimensions is nine), it is difficult for visual data mining. While it is possible to encode several more dimensions into a graph by using various symbols and/or colors, the human perceptual system is not prepared to deal with more than three dimensions simultaneously. Principal component analysis (PCA) is an effective unsupervised method for achieving dimensionality reduction. Using PCA, we can find those orthogonal axes onto which the projections retain maximal variance. Thus a lower dimensional new representation of the set of observed vectors in the space represented by the principal component axes. However, by examining the limitation of PCA, we find that it may not be proper to fulfill our expected role in our feature data structure discovery since we not only have to capture maximal information from the feature data, but also need to cluster the data points to identify the data territory in the space. Another concern is the identification of features among all calculated features, which is responsible for cluster separation and further mass and non-mass classification. The limitation of PCA is that the dimensions with large variances but small cluster separability may play dominant roles in determining the projections and further mislead the dimensionality reduction for cluster separation purpose.

We move the conventional PCA to a direction in which it may serve as a discriminant criterion so that clusters are to be separated and visualized to meet the need of cluster separation. While conventional PCA is an unsupervised method, discriminative principal component analysis (DPCA) is a supervised method that is applied when prior knowledge of class information has been obtained. Based on the class information, a better way of finding directions for cluster separation, however, is to emphasize the inter-cluster separation by using Fisher's scatter matrix instead of total covariance matrix in conventional PCA. This is a discriminative projection searching process,

$$\mathbf{W} = \arg\max_{\mathbf{W}_0}\{Trace(\mathbf{W}_0^{\mathrm{T}}\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{W}_0)\}$$

where $\mathbf{S}_w$ is the within-cluster scatter matrix, $\mathbf{S}_b$ is the between-cluster scatter matrix, and **W** is the optimum projection matrix. This is termed as discriminative principal component analysis (DPCA).

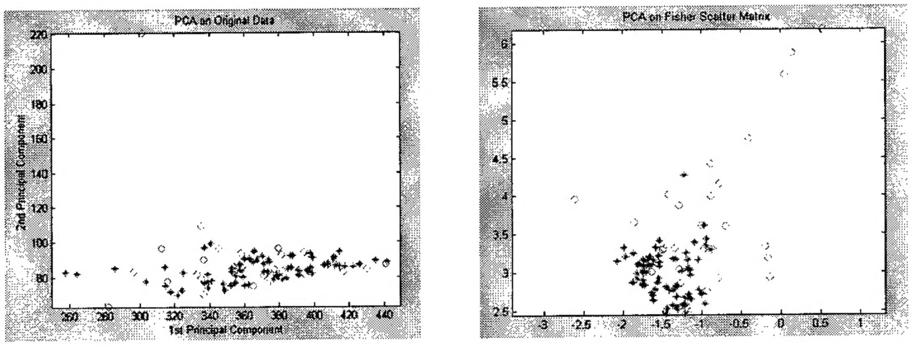


Fig.1. 2 –D projections of feature data using conventional PCA and DPCA, * -- mass and ○ -- non_mass.

From fig. 1, we can see the difference between projections resulting from conventional PCA and discriminative PCA on the effect of cluster separation. In the left figure conventional PCA is applied, mass and non-mass data points are mixed uniformly without revealing any cluster structure, while mass and non-mass data points define clearer distribution structure in the right figure that is resulted from discriminative PCA.

3.1.2 Hierarchical Structure

The According to Cover's theorem on the separability of patterns, when a data set is linearly projected onto a single dimension-reduced subspace, its inherent multi-modal nature may be partially or completely obscured. The revelation of growing volume of high dimensional and multi-modal data set demands a data mining tool differing from conventional data visualization method, which is capable of dealing with high dimensional data set. This motivates our consideration of a hierarchical visualization paradigm involving hierarchical statistical models and visualization space. Comprehensive studies on this issue brought us the possibility of using several complementary visualization subspaces to accomplish this complicated task. In this algorithm, dimensionality reduction and cluster decomposition are two major components. The cluster decomposition permits the use of relatively simple models for each local structure, offering great ease of interpretation as well as many benefits of analytical and computational simplification. On the other hand, dimensionality reduction allows visual explanation of high dimensional data set and less computational demand. We proposed using standard finite normal mixtures (SFNM) and hierarchical visualization spaces for as effective data modeling and visualization. The strategy is that top level model and projection should explain the whole structure of the data set, while lower level models explain the local and internal structure between individual cluster, which may not be obvious in the high level models. With many complementary mixture models and visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet still conveys considerable cluster information. Fig.2. shows an example of a hierarchical visualization tree generated using a set simulated data. The left figure is a top level projection of the data where we can only see two clusters without incorporating color information, the upper right figure is a second level projection that provides different views of two sub-clusters selected in the top level projection. In the second level, we can see two hidden clusters in sub-cluster #2 in the projection differing from the top view, this gives the user opportunities to discover true data structure and makes further partitioning possible.

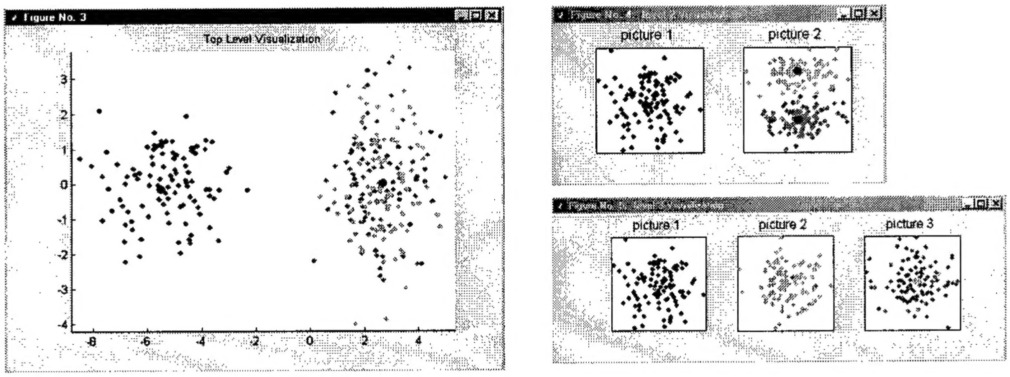


Fig. 2. User Interface of Visual Data Exploration and Mining Tool

3.1.3 User Interaction

User interaction with the algorithm is also an important issue. We have developed a user-friendly graphical interface to facilitate the data visualization purpose, as shown in Fig. 1, which allows the user to select initial centers of the data clusters. Our experience has convincingly indicated a great reduction of both computational complexity and local optimum likelihood. It should be pointed out that although the final SFNM model can be estimated, the pathways of achieving cluster decomposition may be multiple. For example, in this case the user has the flexibility to select only two clusters in the second level and to further split the ``right" cluster, thus to adopt a three-level hierarchy. We believe that this user-driven nature of the current algorithm is also highly appropriate for the visualization context.

3.2 Feature Database Exploration for Feature Selection and Classifier Design

As the primary goal of the visual explanation and mining tool development, we use it to reveal and explain feature database structure for CAD design purpose. We try to make both hidden data patterns and neural network "black box" to be as transparent as possible to users, such as radiologists and patients, through interactive visual explanation.

3.2.1 Feature Selection

We tried to rank and select features that are responsible for differentiating mass and non-mass cases. One of advantages of the work is to reduce the computation load for classifier via reducing the dimension of the feature dataset. Also, the performance of classifier may even be improved if only the features with high discrimination power are used while the non-discriminative features are discarded. Although the simple method of selecting just the best individual feature without considering dimension dependence may fail dramatically, it might still be worthy as a first step. We applied our software to model the dataset with an SFNM distribution. Based on the distribution model, we can perform DPCA to determine the top discriminative principal axes. The nine features were ranked in their discriminative power from high to low: energy, sum entropy, compactness, inertia, sum average, entropy, correlation, difference entropy, and Inverse difference moment. This result is in turn fully used in the classifier design that will be discussed in the following section.

3.2.2 Neural Network Classifier Design

In classifier selection and design, feature database structure is the major guidance we can depend on. All these approaches have the only important goal that is to improve CAD performance in a rational way so that we can explain how we design each component of the CAD system, why such an integrated system works or does not work, and further explain to radiologists to get feedback on the development, the process is fairly transparent to users. Based on the feature ranking, we designed a backpropagation multiple layers network classifier, with one input layer, two hidden layers and one output layer. The number of inputs was reduced from nine to three by using the top three features resulting from the feature ranking. The performance of the classifier has been analyzed using receiver operating characteristics (ROC), also the resulting performance was compared with that of the classifier using all features as inputs. The comparison showed that the three top features together can completely represent the full feature dataset in term of classification, and further more the performance is even better, which is shown in fig. 3. The Az value of the classifier with three inputs is 0.78, for the classifier using all features it is 0.68. Although the results are still not very promising, such a design approach is giving us much more understanding of how the feature database can be used for classifier design. Not only is the success of feature ranking and selection in classifier design reflected in the lowering computational cost through dimension reduction, but also implies that the combination of top rated feature has more discriminative power in the classification.

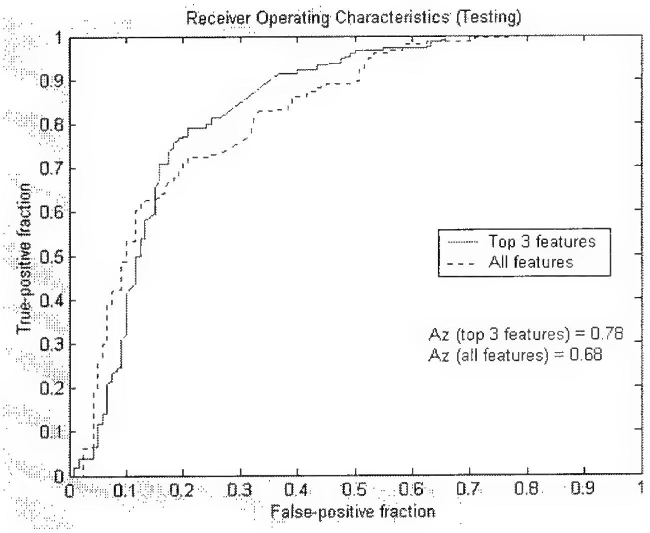


Fig. 3. ROC analysis of neural network classifier

Key Research Accomplishments

- Improving research skill and enhancing fundamental engineering knowledge through book and literature searching and reading under guidance of advisor and mentors.
- Collecting image cases, processing images by computing image features and constructing high dimensional image feature database.
- Developing and improving visual data explanation and mining tool and exploring feature database structure for feature selection and classifier design to make the CAD design processing effective and reasonable.
- Designing classifier and improving classifier performance based on data structure exploration and the study of features.

Reportable Outcomes

- Y. Wang, Z. Wang, L. Luo, S-H. B. Lo and M. T. Freedman, "Computer-Based Decision Support System: Visual Mapping of Featured Database in Computer-Aided Diagnosis", *Proc. Of SPIE, Image Processing,* Vol. 1, No. 24, pp. 136-147, February 2000.
- Feature extraction programs.
- Visual data exploration and mining tool software.
- J. Lu, Y. Wang, Z. Wang, et. al., "Discriminative Mining of Gene Microarray Data", *Proc. Of Neural Networks for Signal Processing,* pp. 23-32, September, 2001.

Conclusions

In this project, we devoted efforts in developing effective feature extraction methods, constructing feature database, developing visual explanation tool for data mining and knowledge discovery, which is both statistically principled and visually effective. This method, as illustrated by the well-planned simulations and pilot applications in computer-aided diagnosis, can be very capable of revealing hidden structure within data. It is important to emphasize that in relation to previous work, one interesting consideration with the present algorithm is that the models are determined by the information theoretic criteria, and this criterion can not only select the most appropriate model structure but also allow a user-driven portfolio as a double check. This approach promotes a self-consistent fitting of the whole tree, so that an automated procedure for generating the hierarchy becomes reality. In addition, since we perform model selection and parameter initialization firstly over the projection space, the computational complexity is greatly reduced in compared to the maximum likelihood estimation in full dimension. Other possible advantages include the determination of data projection by maximizing the separation of clusters, which in turn optimizes the other crucial operations such as model selection and parameter initialization, which help user find hypothesis driven nature of the data projection. Using the visual explanation tool, we tried to discover the feature database structure for feature selection and also classifier design. The performance of the classifier reflected that the feature selection based on feature ranking also makes it possible to reduce dimensionality in classifier design besides in visual data exploration software.

List of Personnel Receiving Pay

1. Zuyi Wang, Department of Electrical Engineering, The Catholic University of America, Washington, DC.
2. Lan Luo, worked in the first year of this project.

References

1. Y. Wang, Z, Wang, L. Luo, S-H. B. Lo and M. T. Freedman, "Computer-based decision support system: visual mapping of featured database in computer-aided diagnosis", *Proc. Of SPIE, Image Processing,* Vol. 1, No. 24, pp. 136-147, February 2000.
2. Y. Wang, L. Luo, M. T. Freedman and S-Y. Kung, "Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization", *IEEE Trans. on Neural Networks,* Vol. 11, No. 3, pp. 625-636, May 2000.
3. C. M. Bishop and M. E. Tipping, ``A hierarchical latent variable model for data visualization," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 20, No. 3, pp. 282-293, March 1998.
4. L. Luo, Y. Wang, and S. Y. Kung, ``Hierarchy of probabilistic principal component subspaces for data mining," *Proc. IEEE Workshop on Neural Nets for Signal Processing*, Wisconsin, August 1999.
5. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

Computer-Based Decision Support System: Visual Mapping of Featured Database in Computer-Aided Diagnosis

Y. Wang[a], Z. Wang[a], L. Luo[a], S-H B. Lo[b], M. T. Freedman[b]

[a]Department of Electrical Engineering and Computer Science
The Catholic University of America, Washington, DC 20064, USA

[b]Department of Radiology and the Lombardi Cancer Center
Georgetown University Medical Center, Washington, DC 20007, USA

ABSTRACT

As a strategic move toward improving the utility of computer-aided diagnosis (CAD) in breast cancer detection, this work aims to develop a computer-based decision support system, through a visual mapping of featured database, to explain the entire decision making process jointly by the computer-encoded knowledge and the user-interaction. The main purpose of the work is twofold: enhance the clinical utility of CAD and provide a mechanism for optimal system design. We adopt a mathematical feature extraction procedure to construct the featured database from the suspicious mass sites localized by the enhanced segmentation. The optimal mapping of the data points is then obtained by learning a hierarchical normal mixtures and associated decision boundaries. A visual explanation of the decision making is further invented through a multivariate data mining and knowledge discovery scheme. In particular, using multiple finite normal mixture models and hierarchical visualization spaces, new strategy is that the top-level model and projection should explain the entire data set, best revealing the presence of clusters and relationships, while lower-level models and projections should display internal structure within individual clusters, such as the presence of subclusters, which might not be apparent in the higher-level models and projections. We demonstrate the principle of the approach on several multimodal numerical data sets, and we then apply the method to the visual explanation in CAD for breast cancer detection from digital mammograms.

1. INTRODUCTION

In order to improve mass detection and classification in clinical screening and/or diagnosis of breast cancers, many sophisticated computer-assisted diagnosis (CAD) systems have been recently developed. Although the clinical roles of the CAD systems may still be debatable, the fundamental role should be complementary to the radiologists' clinical duties or for automated high risk population screening. Literature survey has indicated that (1) most CAD systems are "black" boxes to the users and (2) no working link between "evaluation" and "improvement". This paper addresses the further development of CAD for mass detection based on (1) construction of featured knowledge database; (2) mapping of classified and unclassified data points; and (3) development of a visual exploration and explanation interface.

Although many previously proposed approaches have led to impressive results, several fundamental issues remain unresolved. For example, *Receiver Operating Characteristics* (ROC) analysis can provide an overall performance evaluation, it may not help the improvement of each of the multiple components in CAD system. Furthermore, since the machine observer and human observer may not detect the same set of masses, the "black box" nature of most CAD systems may prevent a natural on-line integration of human intelligence and further upgrade of a CAD system. Our effort is to: (1) provide a visual map of featured database before knowledge encoding component, so to evaluate and improve the pre-processing and signature extraction; (2) based on the map to design an optimal classifier best fitted to this particular database structure for knowledge encoding; and (3) combine the map, the classifier output, raw image, and user interface to explore and explain the whole decision making process by both radiologist and CAD systems.

Further author information: Send correspondence to Y. Wang (E-mail wang@pluto.ee.cua.edu).

In *Medical Imaging 2000: Image Processing*, Kenneth M. Hanson, Editor,
Proceedings of SPIE Vol. 3979 (2000) • 1605-7422/00/$15.00

2. BACKGROUND

As the first step toward understanding multivariate data sets, cluster information reveals insight that may prove useful in knowledge discovery since the growing volume of complex data are often high dimensional, multimodal, and lacking in prior knowledge..[4–6,9] Several new visualization methods have been progressively developed to model and display the contents of the data sets.[4,6–9,11,14] However, although such algorithms can usefully characterize the content of simple data sets, little comprehensive study has been reported that proves adequate in the face of multimodal and high dimensional data sets.[4,9,14] For example, a single projection of the data onto a visualization space may not be able to capture all of the interesting aspects of the data set. This motivates the consideration of a hierarchical visualization paradigm involving hierarchical statistical models and visualization spaces.

Once we explore the possibility of using many complementary visualization subspaces, cluster decomposition and dimensionality reduction are the two major steps. Cluster decomposition permits the use of relatively simple models for each of the local structures, offering greater ease of interpretation as well as the benefits of analytical and computational simplification. On the other hand, dimensionality reduction allows better visual interpretation and less computational demand. Many researchers have recently proposed various methods to improve data visualization.[6,9] The work most closely related to our methodology was reported by Bishop and Tipping in.[4,12] They introduce a hierarchical modeling and visualization algorithm based on a two-dimensional hierarchical mixture of latent variable models, whose parameters are estimated using the expectation-maximization (EM) algorithm.[4,19] The construction of the hierarchical tree proceeds top down in which the cluster decomposition is driven interactively by the user, and optimal projection is determined by maximum likelihood principle.

In this paper, we propose using standard finite normal mixtures (SFNM) and hierarchical visualization spaces for an effective data modeling and visualization. The strategy is that the top-level model and projection should explain the entire data set, best revealing the presence of clusters and relationships, while lower-level models and projections should display internal structure within individual clusters, such as the presence of subclusters, which might not be apparent in the higher-level models and projections. With many complementary mixture models and visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet still conveys considerable cluster information. Based on the concept of combining finite mixture modeling[19] and principal component projection[4,14] to guide cluster decomposition and dimensionality reduction, the particular advantages of our algorithm are:

1. At each level, a probabilistic principle component extraction is performed to project the softly partitioned data set down to a two-dimensional visualization space, leading to an effective dimensionality reduction, allowing effective separation and visualization of local clusters[4,8,15];

2. Learning from the data directly, information theoretic criteria are used to select model structures and estimate its parameter values, where the soft partitioning of the data set results in a standard finite normal mixture distribution best fitted to the data[7,21–25];

3. By alternatively performing principal component projection and finite mixture modeling, a complete hierarchy of complementary projections and refined models can be generated automatically, allowing a new paradigm of knowledge discovery.[4–6,9]

3. THEORY AND METHOD

One of the difficulties inherent in data visualization is the problem of visualizing multi-dimensionality.[4,6,9] When there are more than three variables, it stretches the imagination to visualize their relationships. Fortunately in data set with many variables, groups of variables often form clusters.[13,15,16] Thus, our approach includes two major complementary components: (1) dimensionality reduction by probabilistic principal component projection and (2) cluster decomposition by adaptive soft data clustering.

Assume the data points $\{\mathbf{t}_i\}$ in the data space come from K_0 clusters $\{\boldsymbol{\theta}_{\mathbf{t}1}, ..., \boldsymbol{\theta}_{\mathbf{t}k}, ..., \boldsymbol{\theta}_{\mathbf{t}K_0}\}$, where $\boldsymbol{\theta}_{\mathbf{t}k}$ is the Gaussian kernel parameter vector of cluster k in the model. Recently there has been considerable success in using the SFNM to model the distribution of a multimodal data set,[4,7,10,19,26] such that the data distribution takes a sum of the following general form:

$$p(\mathbf{t}) = \sum_{k=1}^{K_0} \pi_k g(\mathbf{t}|\boldsymbol{\theta}_{\mathbf{t}k}) \tag{1}$$

where π_k is the corresponding mixing proportion, with $0 \leq \pi_k \leq 1$ and $\sum \pi_k = 1$, and g is the Gaussian kernel. The problem of SFNM modeling addresses the combined estimation of regional parameters $(\pi_k, \boldsymbol{\theta}_{\mathbf{t}k})$ and detection of structural parameter K_0 in Eq. (1) based on the observations $\mathbf{t}$. One natural criterion used for estimating the parameter values is to minimize the distance between the SFNM distribution $f(\mathbf{t})$ and the data histogram $f_{\mathbf{t}}$. Suggested by information theory,[19,20] relative entropy (Kullback-Leibler distance) is a suitable measure, given by

$$D(f_{\mathbf{t}}||f) = \sum_{\mathcal{T}} f_{\mathbf{t}}(\mathbf{t}) \log \frac{f_{\mathbf{t}}(\mathbf{t})}{f(\mathbf{t})}. \tag{2}$$

We have previously shown that distance minimization based on (2) is equivalent to the maximum likelihood (ML) estimation under a data independency approximation,[7] and when K_0 is given, the ML estimate of the regional parameters can be obtained using the EM algorithm.[15,19,26]

There are three major problems associated with the current approach. First, when the dimension of the data space is high, the computational complexity of implementing the EM algorithm in $\mathbf{t}$-space is very high. Second, the initialization of the EM algorithm is often heuristically chosen, which may lead to both local optima and computational complexity. Finally, since the number of the local clusters in a particular data set is generally unknown, model selection is a prerequisite. A natural way, with greater practical applicability, to tackle these problems is to introduce user interaction with the system.[4,9] Data mining and knowledge discovery are not processes that can be orchestrated a priori. Training algorithms and expected behavior can be specified, but the actual learning must follow for insight and spontaneous inspiration.[9] For example, by examining plots of principal component space, researchers often develop a deeper understanding of the driving forces that generated the original data, and effortlessly grasp the general characteristics of the data and propose an initial solution.[4,6,9]

Principal component analysis (PCA) is an effective method for achieving dimensionality reduction.[11,12] For a set of observed d-dimensional data vectors $\{\mathbf{t}_i\}$, $i \in \{1, ..., N\}$, the q principal axes $\mathbf{w}_m$, $m \in \{1, ..., q\}$, are those orthogonal axes onto which the retained variance under projection is maximal. It can be shown that the principal axes $\mathbf{w}_m$ are given by the q dominant eigenvectors (i.e., maximal eigenvalues) of the sample covariance matrix $\mathbf{C}_{\mathbf{t}} = \sum_i (\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})^{\mathrm{T}}/N$ such that $\mathbf{C}_{\mathbf{t}}\mathbf{w}_m = \lambda_m \mathbf{w}_m$ and where $\boldsymbol{\mu}_{\mathbf{t}}$ is the sample mean. The vector $\mathbf{x}_i = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_q)$, is thus a q dimensional reduced representation of the observed vector $\mathbf{t}_i$. The advantage of PCA is twofold: the projection onto the principal subspace (1) minimizes the squared reconstruction error[12,15] and (2) maximizes the separation of data clusters.[16] Although the effectiveness of applying PCA in an unsupervised manner is highly data-dependent, our approach has a simple optimal appeal in that if the local clusters are linearly separable in a two- or three-dimensional space, the principal component projections allow best separation of the clusters.[16]

Suppose the data space is d-dimensional. Now consider a two-dimensional projection space $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ together with a linear transformation, that maps the data space to the projection space by $\mathbf{x} = \mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu}_{\mathbf{t}})$ where $\mathbf{W}$ is a $d \times 2$ matrix. For a normal distribution $p(\mathbf{t})$ over the data space, using the rules of probability, a similar reduced dimension probability distribution of the new variables $\{\mathbf{x}_i\}$ in the projection space is obtained from the convolution of the projection model with the true distribution over data space in the form of $f(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{t})p(\mathbf{t})d\mathbf{t}$.[4,12,17] Since the conditional distribution $p(\mathbf{x}|\mathbf{t}) = \delta(\mathbf{x} - \mathbf{W}^{\mathrm{T}}\mathbf{t} + \mathbf{W}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}})$, where $\delta(.)$ is the delta function that $\delta(0) = 1$ and $\delta(\neq 0) = 0$, it can be shown that $f(\mathbf{x})$ is simply defined by the Radon transform of $p(\mathbf{t})$, i.e., $f(\mathbf{x}) = \int p(\mathbf{t})\delta(\mathbf{x} - \mathbf{W}^{\mathrm{T}}\mathbf{t} + \mathbf{W}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}})d\mathbf{t}$.[18] According to the linear superposition property of Radon transform and the projection invariant property of normal distribution, if $p(\mathbf{t})$ is a SFNM distribution, the data distribution in the projection space has a similar reduced dimension form as Eq. (1)

$$f(\mathbf{x}) = \sum_{k=1}^{K_0} \pi_k \int g(\mathbf{t}|\boldsymbol{\theta}_{\mathbf{t}k})\delta(\mathbf{x} - \mathbf{W}^{\mathrm{T}}\mathbf{t} + \mathbf{W}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}})d\mathbf{t} = \sum_{k=1}^{K_0} \pi_k g(\mathbf{x}|\boldsymbol{\theta}_{\mathbf{x}k}). \tag{3}$$

However, because of its global linearity, the application of PCA is necessarily somewhat limited.[12,13] For example, the inherent multimodal nature of the data set may be completely obscured when it is projected onto the lower dimensional principal subspace. Thus, it is important to note that although the cluster structure of the data set may be evident from the higher dimensional plot of the raw data, it is quite conceivable to have the intrinsic cluster structure of the data concealed after a projection in the more general case of high-dimensional data sets.[15] An alternative paradigm is to model multimodal data set with a collection of local linear subspaces through probabilistic principal component analysis as shown in Fig. 1.[12–14] The method is a two-stage procedure: a soft partitioning of the data space followed by estimation of the principal subspace within each partition. For the sake of computational simplicity, it is reasonable to consider the model parameter values being estimated firstly in the projection space and then further fine tuned in the data space.[14]

The association of a SFNM distribution with PCA offers the possibility of being able to visualize complex data structures through a mixture of probabilistic principal component subspaces. By a simple extension of the maximum a posterior for data classification in the standard K_0-ary Bayes hypothesis testing,[15,20] we can obtain a principal component projection along the desired axes onto which a particular portion of the data set is highlighted, by weighting all of the data points in the whole data set with their posterior probabilities belonging to that portion. This involves a soft clustering of the data points in which instead of any given data point being assigned exclusively to one principal component subspace, the responsibility for its generation is shared among all of the subspaces.

Under the SFNM model defined by Eq. (1), the posterior Bayesian probability z_{ik} of a given data point $\mathbf{t}_i$ belonging to cluster k is

$$z_{ik} = \frac{\pi_k g(\mathbf{t}_i|\boldsymbol{\theta}_{\mathbf{t}k})}{p(\mathbf{t}_i)}. \tag{4}$$

where $k = 1, 2, ..., K_0$ and $\sum_k z_{ik} = 1$. These posterior probabilities, together with the computational simplicity of performing PCA (involving no more than finding the top q eigenvectors of the covariance matrix of the data points) make it a good candidate for the linear subspace in the mixture. The q principal components define the local subspace assumed for the multimodal. The contributions of the input to the k subspace are the activities of the weighted data points $\{\mathbf{t}_{ik}\}$ for input cluster k. This can be obtained by $\mathbf{t}_{ik} = z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})$, where $\boldsymbol{\mu}_{\mathbf{t}k}$ is the weighted sample mean of cluster k:

$$\boldsymbol{\mu}_{\mathbf{t}k} = \frac{\sum_i z_{ik}\mathbf{t}_i}{\sum_i z_{ik}}, \qquad \mathbf{C}_{\mathbf{t}k}=\frac{\sum_i z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})^{\mathrm{T}}}{\sum_i z_{ik}} \tag{5}$$

The subspaces for the focused clusters are generated by a localized linear PCA such that $\mathbf{C}_{\mathbf{t}k}\mathbf{w}_{mk} = \lambda_{mk}\mathbf{w}_{mk}$. It is important to understand that each component in Eq. (1) now corresponds to an independent subspace model with parameters $\boldsymbol{\theta}_{\mathbf{x}k}$ and $\mathbf{W}_k$, where $\mathbf{W}_k = (\mathbf{w}_{1k}, \mathbf{w}_{2k}, ..., \mathbf{w}_{qk})$. More precisely, consider the vector $\mathbf{x}_{ik} = z_{ik}\mathbf{W}_k^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})$ to be a q dimensional reduced representation of k-cluster focused vector $\mathbf{t}_{ik}$, the corresponding probability distribution is defined by

$$g(\mathbf{x}|\mathbf{W}_k, \boldsymbol{\theta}_{\mathbf{x}k}) = \int g(\mathbf{t}|\boldsymbol{\theta}_{\mathbf{t}k})\delta(\mathbf{x} - \mathbf{W}_k^{\mathrm{T}}\mathbf{t} + \mathbf{W}_k^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}k})d\mathbf{t} \tag{6}$$

where the data mapping by $\mathbf{W}_k$ leads to an independent Radon transform. To interpret the corresponding set of visualization subspaces, it may be useful to plot all of the data points on every plot. For this, we may create a k-cluster focused projection in k-subspace by plotting the vector $\mathbf{x}_{ik}$, or display the density of "gray-level" in proportion to the contribution which each point has for k-subspace with $h[\mathbf{W}_k^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})] = z_{ik}$.

An important issue concerning unsupervised cluster decomposition is the detection of the structural parameter K_0, called model selection.[7,14,15,19,25] This is indeed particularly critical in real-world applications where the structure of the data patterns may be arbitrarily complex.[5] We propose to use two information theoretic criteria, i.e., the Akaike information criterion (AIC)[21] and minimum description length (MDL),[22] to guide model selection. The major thrust of this approach has been the formulation of a model fitting procedure in which an optimal model is selected from the several competing candidates such that the selected model best fits the observed data, under Jaynes' minimax entropy principle stated as *"the parameters in a model which determine the value of the maximum entropy should be assigned values which minimize the maximum entropy"*.[23,24] For example, AIC tries to reformulate the problem explicitly as an *approximation* of the true structure by the model, implying that AIC will select the model that gives the minimum value defined by

$$\mathrm{AIC}(K_a) = -2\log(\mathcal{L}_{ML}) + 2K_a \tag{7}$$

where $\mathcal{L}_{ML}$ is the maximum likelihood of the model and K_a is the number of free adjustable parameters in the model. From a quite different point of view, MDL reformulates the problem explicitly as an information coding problem in which the best model fit is measured such that it assigns high probabilities to the observed data while at the same time the model itself is not too complex to describe.[22] A model is selected by minimizing the total description length defined by

$$\mathrm{MDL}(K_a) = -\log(\mathcal{L}_{ML}) + 0.5K_a \log N. \tag{8}$$

where the penalty term in MDL takes into account the number of observations. It should be pointed out that when the cluster separability is poor, the performance of these two information theoretic criteria may not be reliable.[21,25]

As discussed above, the SFNM model identification is first performed over x-space. However, a mapping from t-space to x-space may have the intrinsic cluster structure concealed, leading to an incorrect correspondence between Eq. (1) and Eq. (3). We now extend the mixture representation of Eq. (1) to form a hierarchical mixture model generally enough to be applicable to mixtures of any parametric density model. Based on the discussion of a two-level system consisting of a single Radon transform at the top level and a mixture of K_0 normal distributions at the second level, we can reformulate the hierarchy to a third level by associating a group $\mathcal{G}_k$ of SFNM models with each model k in the second level, given by

$$p(\mathbf{t}) = \sum_{k=1}^{K_0} \pi_k \sum_{j=1}^{L_{k,0}} \pi_{j|k} g(\mathbf{t}|\boldsymbol{\theta}_{\mathbf{t}(k,j)}) \tag{9}$$

where $\pi_{j|k}$ again correspond to a set of mixing proportions, one for each k, with $\sum_j \pi_{j|k} = 1$. The formation of the hierarchy is guided by the model selection over x-subspaces, where each level of the hierarchy corresponds to a generic model, with lower levels giving more focused and interpretable representations. Once again each component in Eq. (9) now corresponds to an independent subspace model with Radon transform $g(\mathbf{x}|\boldsymbol{\theta}_{\mathbf{x}(k,j)}) = \int g(\mathbf{t}|\boldsymbol{\theta}_{\mathbf{t}(k,j)})\delta(\mathbf{x} - \mathbf{W}_{(k,j)}^{\mathrm{T}}\mathbf{t} + \mathbf{W}_{(k,j)}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}(k,j)})dt$.

4. ALGORITHMS

Based on the theory behind hierarchical mixtures of probabilistic principal component subspaces we have discussed above, we now present the description of our algorithm involving major steps of the visual hierarchy construction. Although the tree structure of the hierarchy may be empirically defined,[4,12] a more interesting effort, is to build the tree *automatically and interactively.* Guided by the two information theoretic criteria, our algorithm progressively proceeds by fitting a series of submodels to the clusters of the data set, in which model order is selected automatically and algorithm initialization is driven interactively. A schematic summary of the algorithm is as follows:

1. Project the data set onto a single x-space, in which $\mathbf{W}$ is determined from the sample covariance matrix $\mathbf{C}_\mathbf{t}$ by fitting a single Gaussian model to the data set over t-space.

2. Learn $f(\mathbf{x})$ for $K = K_{MIN}, ..., K_{MAX}$, in which the values of π_k and $\boldsymbol{\theta}_{\mathbf{x}k}$ are initialized by the user and estimated by the EM algorithm over x-space.

3. Calculate the values of AIC and MDL for $K = K_{MIN}, ..., K_{MAX}$, and select a model with K_0 which corresponds to the minimum of AIC and MDL. The model parameters obtained in x-space will be used to initialize the model parameters in t-space for the learning in step 4.

4. Learn $f(\mathbf{t})$ with K_0, in which the values of π_k, z_{ik}, $\boldsymbol{\mu}_{\mathbf{t}k}$, and $\mathbf{C}_{\mathbf{t}k}$, are fine tuned by the EM algorithm over t-space..

5. Determine $\mathbf{W}_k$ from $\mathbf{t}_{ik}$ or $\mathbf{C}_{\mathbf{t}k}$, and plot $\mathbf{x}_{ik}$ or $h[\mathbf{W}_k^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})]$ onto x-subspaces at the second level for visual evaluation, for $k = 1, 2, ..., K_0$.

6. Learn $\mathcal{G}_k(\mathbf{t})$ by repeating steps $2 - 4$ and construct x-subspaces at the third level by repeating step 5, for $k = 1, 2, ..., K_0$.

7. Complete the whole hierarchy under the information theoretic criteria, and plot all x-subspaces for visual exploration and explanation.

Our algorithm begins by determining $\mathbf{W}$ for the top level projection. For low dimensional data sets, we directly evaluate the covariance matrix $\mathbf{C_t}$ to find $\mathbf{W}$.[13,15] For high dimensional cases, since only the top two eigenvectors of the covariance matrix of the data points are of the interest, it may be computationally more efficient to apply our previously developed APEX neural networks[8] to find $\mathbf{W}$ directly from the data points $\mathbf{t}_i$ (Step 1). On the basis of this single x-space, given a fixed K, the user then selects (K_{MIN}, K_{MAX}) and points $\boldsymbol{\mu}_{\mathbf{x}k}$ on the plot corresponding to the centers of apparent clusters. The EM algorithm can be applied to allow a SFNM (Eq. (3)) to be fitted to the projected data through the following two-stage[19,26] form:

E-Step

$$z_{ik}^{(n)} = \frac{\pi_k^{(n)} g(\mathbf{x}_i|\boldsymbol{\theta}_{\mathbf{x}k}^{(n)})}{f(\mathbf{x}_i|\pi_k^{(n)}, \boldsymbol{\theta}_{\mathbf{x}k}^{(n)})} \tag{10}$$

M-Step

$$\pi_k^{(n+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{ik}^{(n)}, \quad \boldsymbol{\mu}_{\mathbf{x}k}^{(n+1)} = \frac{\sum_{i=1}^{N} z_{ik}^{(n)}\mathbf{x}_i}{\sum_{i=1}^{N} z_{ik}^{(n)}}, \quad \mathbf{C}_{\mathbf{x}k}^{(n+1)} = \frac{\sum_{i=1}^{N} z_{ik}^{(n)}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}k}^{(n)})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}k}^{(n)})^{\mathrm{T}}}{\sum_{i=1}^{N} z_{ik}^{(n)}} \tag{11}$$

where at each complete cycle of the algorithm, we first use "old" set of parameter values to determine the posterior probabilities $z_{ik}^{(n)}$ using Eq. (10). These posterior probabilities are then used to obtain "new" values $\pi_k^{(n+1)}$, $\boldsymbol{\mu}_{\mathbf{x}k}^{(n+1)}$, and $\mathbf{C}_{\mathbf{x}k}^{(n+1)}$ using Eqs.(11). The algorithm cycles back and forth until the value of relative entropy (Eq. (2)) reaches its minimum (Step-2). It can be shown that, at each stage of the EM algorithm, the relative entropy decreases unless it is already at a local minimum.[19] The model selection procedure will then determine the optimal number K_0 of models to fit at the next level down using the two information theoretic criteria, where $K_a = 6K_0 - 1$ including $2K_0$ means, $2K_0$ variances, K_0 correlation coefficients, and $K_0 - 1$ mixing factors (Step 3). The resulting points $\boldsymbol{\mu}_{\mathbf{t}k}^{(0)}$ in data space, obtained by $\boldsymbol{\mu}_{\mathbf{t}k}^{(0)} = \mathbf{W}\boldsymbol{\mu}_{\mathbf{x}k}^{(\infty)} + \boldsymbol{\mu}_{\mathbf{t}}$, are then used as the initial means of the respective submodels. Since the mixing proportions π_k are projection-invariant, we simply assign a 2×2 unit matrix to the remaining parameters of the covariance matrix $\mathbf{C}_{\mathbf{t}k}$. Once again the EM algorithm can be applied to allow a SFNM (Eq. (1)) with K_0 submodels to be fitted to the data over t-space. In order to obviate the need to store all the incoming observations, and change the parameters immediately after each data point, it may be computationally more efficient to apply our previously developed probabilistic self-organizing map (PSOM), an incremental EM algorithm,[7] to estimate $p(\mathbf{t})$.

With a soft partitioning of the data set using the PSOM, data points will now effectively belong to more than one cluster at any given level. Thus, the effective input values are $\mathbf{t}_{ik} = z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})$ for an independent visualization subspace k in the hierarchy. We then extend our APEX algorithm to a probabilistic version, i.e., PAPEX,[8,27] to determine $\mathbf{W}_k$, summarized as follows (Step 4).

1. Initialize the feedforward weight vector $\mathbf{w}_{mk}$ for $m = 1, 2$, and the feedback weight vector a_k to small random values at time $i = 1$. Assign a small positive value to the learning rate parameter η.

2. Set $m = 1$, and for $i = 1, 2, \ldots$, compute

$$y_{1k}(i) = \mathbf{w}_{1k}^{\mathrm{T}}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}), \quad \mathbf{w}_{1k}(i+1) = \mathbf{w}_{1k}(i) + \eta[y_{1k}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}) - y_{1k}^2(i)\mathbf{w}_{1k}(i)] \tag{12}$$

   For large i we have $\mathbf{w}_{1k}(i) \longrightarrow \mathbf{w}_{1k}$, where $\mathbf{w}_{1k}$ is the eigenvector associated with the largest eigenvalue of the covariance matrix $\mathbf{C}_k$.

3. Set $m = 2$, and for $i = 1, 2, \ldots$, compute

$$y_{2k}(i) = \mathbf{w}_{2k}^{\mathrm{T}}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}) + a_k(i) y_{1k}(i), \quad \mathbf{w}_{2k}(i+1) = \mathbf{w}_{2k}(i) + \eta[y_{2k}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}) - y_{2k}^2(i)\mathbf{w}_{2k}(i)] \tag{13}$$

$$a_k(i+1) = a_k(i) - \eta[y_{2k}(i) y_{1k}(i) + y_{2k}^2(i) a_k(i)] \tag{14}$$

   For large i we have $\mathbf{w}_{2k}(i) \longrightarrow \mathbf{w}_{2k}$, where $\mathbf{w}_{2k}$ is the eigenvector associated with the second largest eigenvalue of the covariance matrix $\mathbf{C}_k$.

Having determined principal axes $\mathbf{W}_k$ of the mixture model at the second level, we will construct the visualization subspaces by plotting each data point $\mathbf{t}_i$ at the corresponding $\mathbf{x}_{ik}$. Thus if one particular point takes most of the contribution for a particular component, then that point will effectively be visible only on the corresponding subspace (Step 5).

Determination of the parameters of the models at the third level can again be viewed as a two-step estimation problem, in which further split of the models at the second level is determined within each of the subspaces over x-space, and then the parameters of the selected models are fine tuned over t-space. Similarly, the resulting model estimated over x-space are then used to initialize the means of the respective submodels over t-space. The corresponding $\mathcal{G}_k(\mathbf{t})$ can again be estimated using the EM or PSOM algorithm[7,19,26] to allow a SFNM distribution with $L_{k,0}$ submodels to be fitted to the data. In the E-step, the posterior probability that data point $\mathbf{t}_i$ belongs to submodel j is given by

$$z_{i(k,j)} = z_{ik} z_{ij|k} = z_{ik} \frac{\pi_{j|k} g(\mathbf{t}_i|\boldsymbol{\theta}_{\mathbf{t}j|k})}{\mathcal{G}(\mathbf{t}_i|\boldsymbol{\theta}_{\mathbf{t}k})}. \tag{15}$$

where z_{ik} are constants estimated from the second level of the hierarchy. The corresponding M-step includes

$$\pi_{j|k} = \frac{\sum_{i=1}^N z_{i(k,j)}}{\sum_{i=1}^N z_{ik}}, \quad \boldsymbol{\mu}_{\mathbf{t}(k,j)} = \frac{\sum_{i=1}^N z_{i(k,j)} \mathbf{t}_i}{\sum_{i=1}^N z_{i(k,j)}}, \quad \mathbf{C}_{\mathbf{t}(k,j)} = \frac{\sum_{i=1}^N z_{i(k,j)} (\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})^{\mathrm{T}}}{\sum_{i=1}^N z_{i(k,j)}}. \tag{16}$$

With the resulting $z_{i(k,j)}$ in t-space, we can apply the PAPEX algorithm to estimate $\mathbf{W}_{(k,j)}$, in which the effective input values are expressed by $\mathbf{t}_{i(k,j)} = z_{i(k,j)}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})$. The next level visualization subspace is generated by plotting each data point $\mathbf{t}_i$ at the corresponding $\mathbf{x}_{i(k,j)} = z_{i(k,j)} \mathbf{W}_{(k,j)}^{\mathrm{T}} (\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})$ in (k,j)-subspace (Step 6).

The construction of the entire tree structure hierarchy is automatically completed when no further data split is recommended by the information theoretic criteria in all of the parent subspaces (Step 7).

5. ILLUSTRATION AND APPLICATION

We first illustrate the application of our algorithm to a simple synthetic data set. Fig. 1 (a) shows a data set consisting of 450 data points generated from a mixture of three Gaussians in three-dimensional space. Each Gaussian is relatively flat (has small variance) in one dimension. Two of these pancake-like clusters are closely spaced, while the third is well separated from the first two. The dimensionality of this data set has been chosen to illustrate the basic principle of the approach. The global view of the raw data over t-space clearly suggests the presence of three distinct clusters within the data.

To explore the data characteristics, we first perform a single global PCA to project each data point onto a single x-space (top level), shown in Fig. 1 (b). Both the user inspection and the two information theoretic criteria have clearly suggested the presence of two distinct clusters within the projected data set. Based on a soft clustering of the data points, we then apply PAPEX to both clusters and generate the two corresponding independent cluster-focused subspaces (second level), as shown in Fig. 1 (c). Not to our surprise, the two information theoretic criteria have suggested a further split of cluster 2 but not of cluster 1. Once again by performing three independent PAPEX, the final cluster decomposition through the cluster-focused subspaces (third level) is completed shown in Fig. 1 (d).

With this three-level hierarchical data exploration, the capable nature of the approach is evident as the interim two subspaces (second level) only attempt to highlight the data points which have already been modeled by their immediate ancestor (top level). Indeed, the model fitting procedure has successfully discovered all three data clusters. The original data clusters have been individually colored, and it can be seen that the red, yellow, and blue data points have been well separated and highlighted in the third level subspaces.

As an example of a more complex problem, we consider a data set arising from a mixture of three closely spaced Gaussians consisting of 300 data points, shown in Fig. 2 (a). Once again the original data clusters have been individually colored. We first apply APEX to extract the global principal axis, indicated by the black line in Fig. 2 (a). The two information theoretic criteria have suggested the presence of three distinct clusters, where the user then selects three initial cluster centers and the EM/PSOM algorithm is applied to perform a soft clustering of the data points. This leads to a mixture of three independent probabilistic principal component subspaces whose principal axes are separately extracted, indicated by the yellow lines in Fig. 2 (a). The contributions of each data point to these subspaces, in terms of its "gray-level" $h[\mathbf{t}_i] = z_{ik}$, are displayed over t-space in Fig. 2 (b).

Since the model selection and algorithm initialization are performed over x-space with user's interaction, it may be helpful to investigate the visual effectiveness of dimensionality reduction using the probabilistic principal component projections.[4,9] Based on the estimated $\mathbf{W}_k$, we have constructed each of the cluster-focused subspaces using both "data graphics" (e.g., in terms of $\mathbf{x}_{ik} = z_{ik}\mathbf{W}_k^T(\mathbf{t}_i - \mu_{tk})$) and "data image" (e.g., in terms of $h[\mathbf{W}_k^T(\mathbf{t}_i - \mu_{tk})] = z_{ik}$) techniques. As a more overlapped case, Fig. 2 (c-d) present the plots of "data graphics" and "data image" from the data set, where "data graphics" emphasizes the contribution of a particular data point to that particular subspace concerning its geometric distance to the center of the cluster, while "data image" emphasizes the effectiveness of a data point reflecting its global appearance. It can be seen that the plot of each cluster is clean and well-shaped.

In order to quantitatively evaluate the effectiveness of our approach with user interactions,[9] we apply our algorithm to a synthesized testing data set given in Fig. 3 (up-left). Using the APEX algorithm we accurately estimate the top global principal axis, indicated by the back line. By projecting the data points onto a two-dimensional x-space, all three data clusters are visible. This plot indicates that although the second advantage of PCA forementioned is highly data-dependent, when the data clusters are linearly separable in a projection space, the principal component projections allow effective separation of the clusters.[16] We then apply the two information theoretic criteria to examine this plots. In this case, we set $K_{MIN} = 1$ and $K_{MAX} = 5$. The minima of both AIC and MDL have clearly suggested a three-cluster data structure, as given by the curve in Fig. 3 (third block in the second row). Thus a two-level SFNM model may be sufficient. We then conduct two experiments to assess the performance of our algorithm. Since all the model parameters are known in this case, the true top principal axes of the data clusters have been individually calculated. First, we compare the estimated top principal axes of the data clusters using our algorithm with the corresponding true top principal axes. From the down-right block in Fig. 3, it can be seen that the two sets of the top principal axes are perfectly matched (blue lines). Second, we use the global relative entropy (GRE) between the data histogram and the estimated SFNM model to measure the goodness of model fitting. The numerical result through our experiments indicates a very good performance with a GRE value of 0.008 nats.

User interaction with the algorithm is an important issue. We have developed a user-friendly graphical interface to facilitate the data visualization purpose, as shown in Fig. 3. By allowing the user to select the initial centers of the data clusters demonstrated in Fig. 3, our experience has convincingly indicated a great reduction of both computational complexity and local optimum likelihood. For example, compared to the results of model selection reported by Akaike[21] and Wax,[25] the curves of the AIC and MDL generated by our algorithm are much more consistent and smooth, and user-initialized computation is five times (in average) faster than the random trials. It should be pointed out that although the final SFNM model can be estimated, the pathways of achieving cluster decomposition may be multiple. For example, in this case the user has the flexibility to select only two clusters in the second level and to further split the "right" cluster, thus to adopt a three-level hierarchy. We believe that this user-driven nature of the current algorithm is also highly appropriate for the visualization context.[4,14]

Since a more convincing example should involve more clusters with multiple levels, we have also applied our algorithm to the same data set used by Bishop&Tipping,[4] shown in Fig. 4 (a). This data set arises from a noninvasive monitoring system used to determine the quantity of oil in a multiphase pipeline containing a mixture of oil, water, and gas.[4] The experiment gives 12 diagnostic measurements in total. Our interim goal is to visualize the structure of the data in the original 12-dimensional space. A data set consisting of 1,000 points is obtained synthetically and the data is expected to have an intrinsic dimensionality of two corresponding to the two dominant components (e.g., oil and water). However, the presence of different flow configurations leads to numerous distinct clusters. We then apply our algorithm to perform a cluster discovery. Results from partially fitting the oil flow data using a three-level hierarchical model are given in Fig. 4. It should be pointed out that since the "right" answer to this real-world data set is not available, we are not able to validate this new result. However, we believe that this example has clearly been highly successful, note how the selected single cluster (number 2) in the top-level plot, is discovered to be two quite separated clusters at the second level.

As a final example, we consider the visual explanation in computer-aided diagnosis (CAD) for breast cancer detection. As a step toward improving the performance of CAD system, we have put considerable efforts to conduct various studies and develop reliable image enhancement and lesion segmentation techniques.[7] More precisely, we try to make both the hidden data patterns and the neural network "black box" to be as transparent as possible to the user (e.g., radiologists and patients) through interactive visual explanation. The clinical goal is to eliminate the false positive sites that correspond to normal dense tissues with mass-like appearances through featured discrimination. We adopt a mathematical feature extraction procedure to construct our database from all the suspicious mass sites localized by the enhanced segmentation.[7] The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a probabilistic modular neural network is developed to carry out both soft and hard clustering.[7] The joint histogram of the featured database extracted from true and false mass regions are investigated and the features that can better separate the true and false mass sites are selected.[7] Our experience has suggested that three imagery features, i.e., site area, compactness, and difference entropy, were having good discrimination and reliability properties.

We then use our previously developed algorithm[7] to distinguish the true masses from false masses based on the features extracted from the suspected regions. 150 mammograms were selected from the mammogram database. Each mammogram contained at least one mass case of varying size and location. The areas of suspicious masses were identified following the proposed procedure with biopsy proven results. In a typical experiment, we have selected a three-dimensional feature space consisting of compactness I, compactness II, and difference entropy. It should be noticed that the feature vector can easily extend to higher dimensionality. A training feature vector set was constructed from 50 true mass ROIs and 50 false mass ROIs, where ROI stands for *region of the interest*. In addition to the decision boundaries recommended by the computer algorithms, a visual explanation interface has also been integrated with hierarchical projections. Fig. 5 (a) shows the database map selection with compactness definition I and difference entropy. Fig. 5 (b) shows the database map selection with compactness definition II and difference entropy. Our experience has suggested that the recognition rate with compactness I are more reliable than that with compactness II.

We have conducted a preliminary study to evaluate the performance of the algorithms in real case detection, in which 6 – 15 suspected masses per mammogram were detected and required further clinical decision making. We found that the proposed visual explanation approach, together with CAD system, can reduce the number of suspicious masses with a sensitivity of 84% at a specificity of 82% (1.6 false positive findings per mammogram) based on the database containing 46 mammograms (23 of them have biopsy proven masses). Fig. 6 shows a representative mass detection result on one mammogram with a stellate mass, indicated by the arrow in Fig. 6 (a). After appropriate feature extraction, ten sites with brightest intensity were selected, shown in Fig. 6 (b). The featured vectors of these candidates were submitted against the estimated "probability cloud" for visual explanation as a decision support, together with the opinion recommended by our CAD system. The final results indicated that the stellate mass lesion was correctly detected, confirmed by our experience radiologists, shown in Fig. 6 (c). It should be pointed out that in this real-world application, a higher recognition rate may be controlled by the domain experts in balancing the trade-off between the *false positive* and *false negative* rates.[7]

6. DISCUSSION

We have presented a novel approach to visual explanation for data mining and knowledge discovery, which is both statistically principled and visually effective. This method, as illustrated by the well-planned simulations and pilot applications in computer-aided diagnosis, can be very capable of revealing hidden structure within data. It is important to emphasize that in relation to previous work,[4,11–13] one interesting consideration with the present algorithm is that the models are determined by the information theoretic criteria, and this criterion can not only select the most appropriate model structure but also allow an user-driven portfolio as a double check. This approach promotes a self-consistent fitting of the whole tree, so that an automated procedure for generating the hierarchy becomes reality.[4] In addition, since we perform model selection and parameter initialization firstly over the projection space, the computational complexity is greatly reduced in compared to the maximum likelihood estimation in full dimension. Our case study of a seven dimensional data set has indicated at least a 50% reduction of the computational time. Other possible advantages include the determination of data projection by maximum the separation of clusters which in turn optimizes the other crucial operations such as model selection and parameter initialization,[16] and data rendering algorithms which permit user or hypothesis driven nature of the data projection.[14]

Acknowledgment

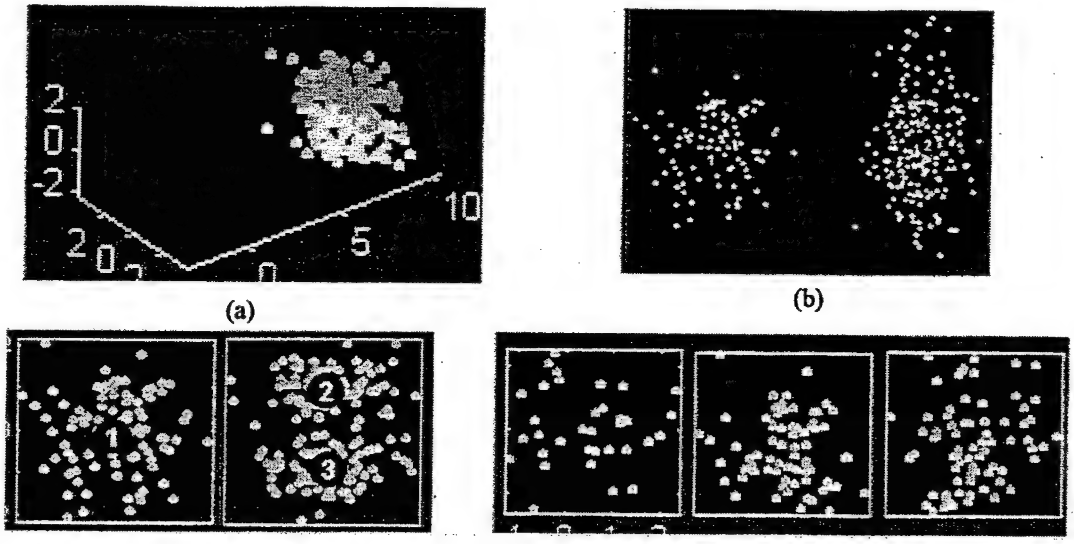
This work is supported by the Department of Defense under Grants DAMD17-98-1-8045 and DAMD17-98-1-8044.

REFERENCES

1. S. M. Lai, X. Li, and W. F. Bischof, "On Techniques for Detecting Circumscribed Masses in Mammograms," *IEEE Trans. on Med. Imaging*, Vol. 8, No. 4, pp. 377-386, 1989.

2. H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Alder, M. M. Goodsitt, and N. Petrick, "Computer-Aided Classification of Mammographic Masses and Normal Tissue: Linear Discriminant Analysis in Texture Feature Space," *Phys. Med. Biol.*, Vol. 40, pp. 857-876, 1995.
3. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. on Sys., Man, and Cyber.*, Vol. SMC-3, No. 6, pp. 610-621, Nov. 1973.
4. C. M. Bishop and M. E. Tipping, "A hierarchical latent variable model for data visualization," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 20, No. 3, pp. 282-293, March 1998.
5. T. R. Golub, D. K. Slonim, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, Vol. 286, pp. 531-537, October 1999.
6. E. R. Tufte, *Visual Explanation: Images and Quantities, Evidence and Narrative*, Graphics Press, Cheshire 1996.
7. Y. Wang, S. H. Lin, H. Li, and S, Y. Kung, "Data mapping by probabilistic modular networks and information theoretic criteria," *IEEE Trans. Signal Processing*, Vol. 46, No.12, pp. 3378-3397, December 1998.
8. S. Y. Kung, *Principal Component Neural Networks*, New York: John Wiley, 1996.
9. G. M. Nielson, "Challenges in visualization research," *IEEE Trans. Visualization and Computer Graphics*, vol. 2, no. 2, pp. 97-99, June 1996.
10. M. I. Jordan and R. A. Jacobs, "Hierarchical mixture of experts and the EM algorithm," *Neural Computation*, Vol. 6, pp. 181-214, 1994.
11. N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, Vol. 9, No. 7, pp. 1493-1516, 1997.
12. M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, Vol. 11, pp. 443-482, 1999.
13. G. E. Hinton, P. Dayan, and M. Revow, "Modeling the manifolds of images of handwritten digits," *IEEE Trans. Neural Net.*, Vol. 8, No. 1, pp. 65-74, January 1997.
14. L. Luo, Y. Wang, and S. Y. Kung, "Hierarchy of probabilistic principal component subspaces for data mining," *Proc. IEEE Workshop on Neural Nets for Signal Processing*, Wisconsin, August 1999.
15. S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, Inc., Upper Saddle River, Ney Jersey, 1999.
16. K. Etemad and R. Chellappa, "Separability-based multiscale basis selection and feature extraction for signal and image classification," *IEEE Trans. Image Processing*, Vol. 7, No. 10, October 1998.
17. R. Gray and L. Davisson, *Random Processes–A Mathematical Approach for Engineers*, Englewood Cliffs, NJ: Prentice-Hall, Inc. 1986.
18. R. N. Bracewell, *Two-Dimensional Imaging*, Prentice-Hall, Inc., 1995.
19. D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical analysis of finite mixture distributions.* New York: John Wiley, 1985.
20. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.
21. H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, Vol. 19, No. 6, pp. 716-723, 1974.
22. J. Rissanen, "Modeling by shortest data description," *Automat.*, Vol. 14, pp. 465-471, 1978.
23. E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, Vol. 106, No. 4, pp. 620-630/171-190, May 1957.
24. J. Rissanen, "Minimax entropy estimation of models for vector processes," *System Identification: Advances and Case Studies*, pp. 97-117, Academic Press, 1987.
25. M. Wax and T. Kailath, "Detection of Signals by Information Theoretic Criteria," *IEEE Trans. Acoust. Speech, Signal Processing,* Vol. 33, No. 2, April 1985.
26. L. I. Perlovsky and M. M. McManus, "Maximum likelihood neural networks for sensor fusion and adaptive classification," *Neural Networks*, Vol. 4, pp. 89-102, 1991.
27. S. Y. Kung, K. I. Diamantaras, and J. S. Taur, "Adaptive Principal Component Extraction (APEX) and Applications," *IEEE Trans. Signal Processing*, Vol. 42, No. 5, pp. 1202-1217, May 1994.

Figure 1



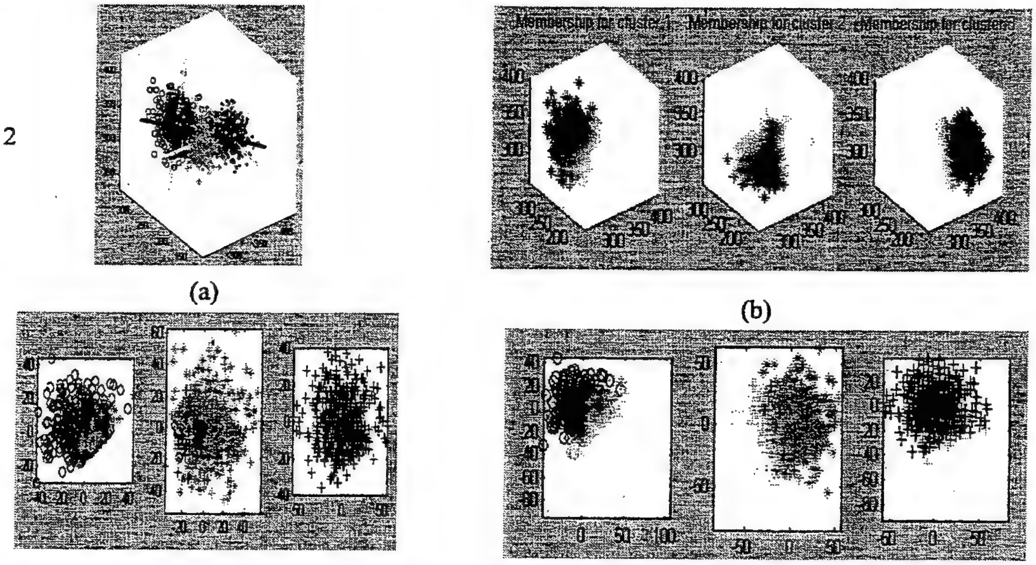
(a)

(b)

(c)

(d)

Figure 2



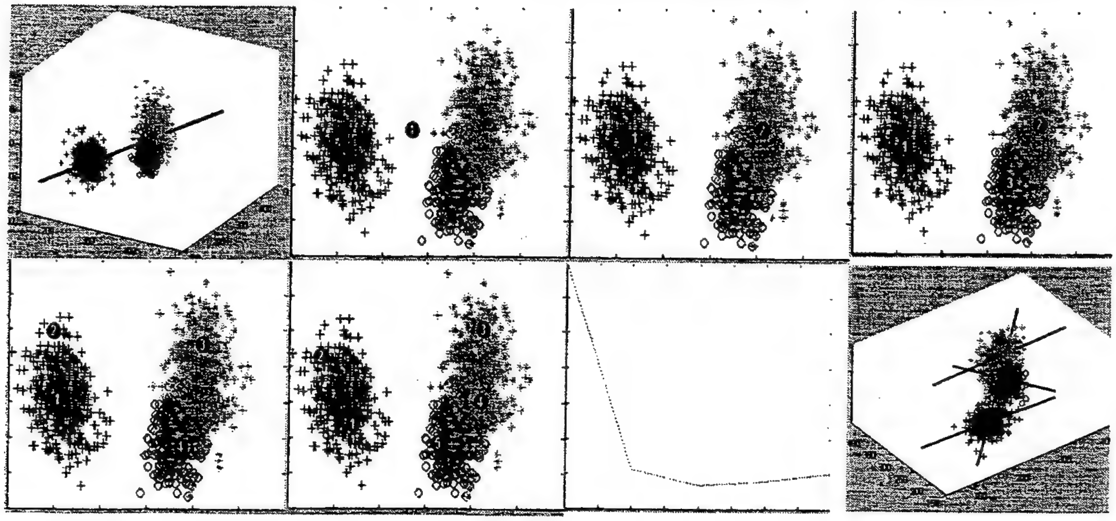
(a)

(b)

(c)

(d)

Figure 3.



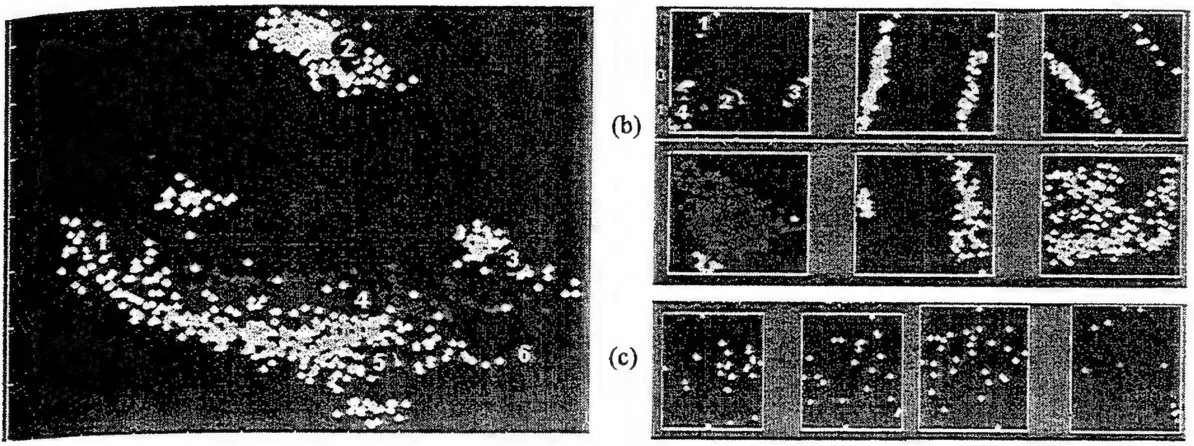


Figure 4. (a)

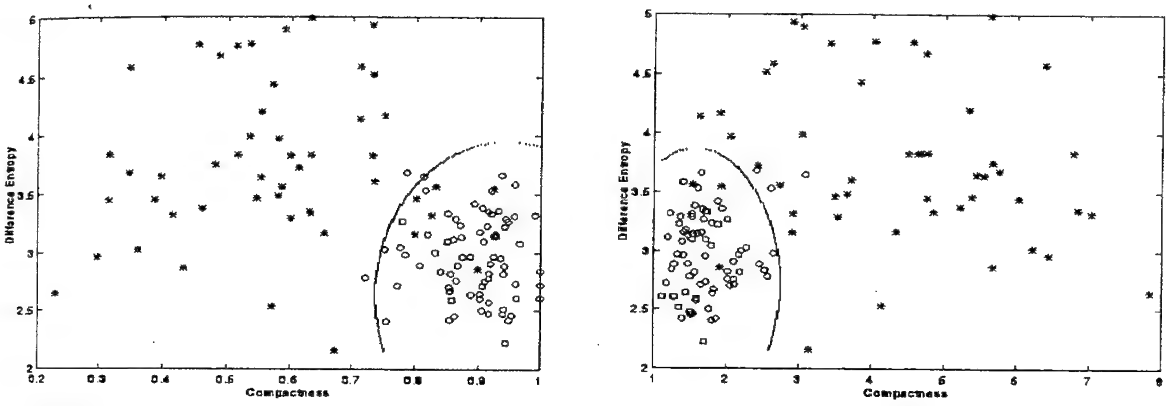
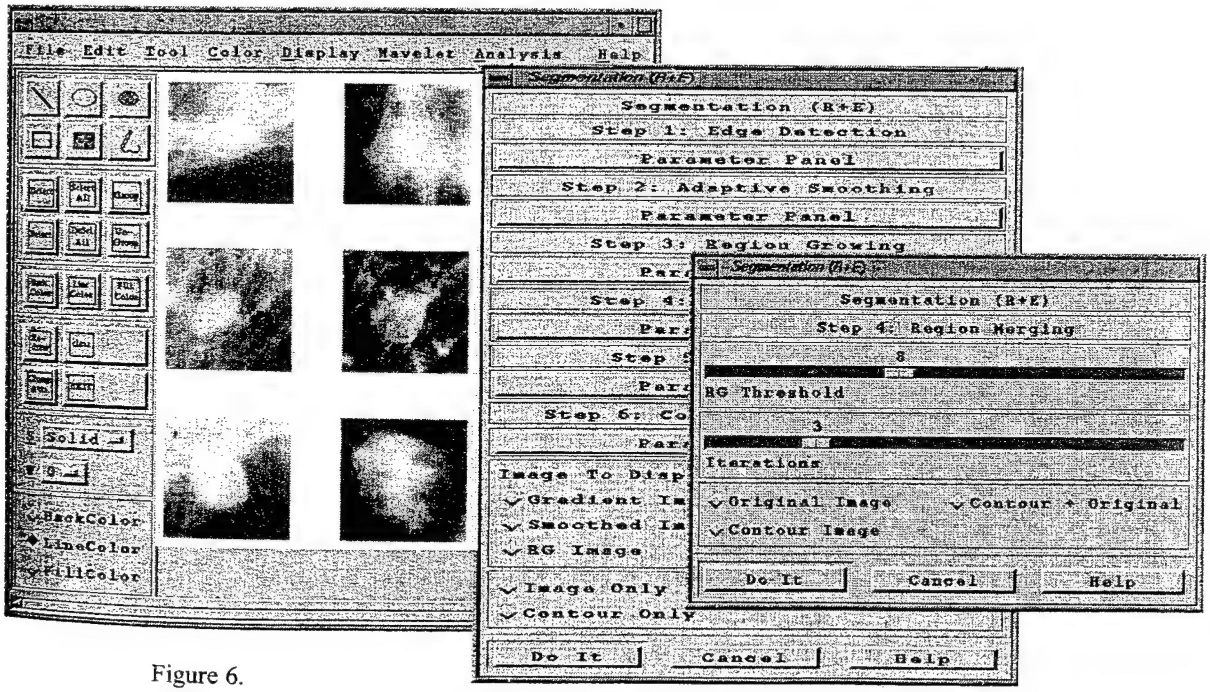


Figure 5. (a) (b)



Segmentation (R+E)

Step 1: Edge Detection

Parameter Panel

Step 2: Adaptive Smoothing

Parameter Panel

Step 3: Region Growing

Step 4:

Step 6:

Image To Display

Gradient Im

Smoothed Im

BG Image

Image Only

Contour Only

Segmentation (R+E)

Step 4: Region Merging

8

RG Threshold

3

Iterations

Original Image

Contour + Original

Contour Image

Figure 6.

REFERENCES

[1] Ashburner, J., & Friston, K.J. (1997). Multimodal Image Coregistration and Partitioning - a Unified Framework. *NeuroImage*, **6**, pp. 209 - 217.

[2] Single subject epoch (block) auditory fMRI activation data. This experiment was conducted by Geriant Rees under the direction of Karl Friston and the FIL methods group. Freely available on `ftp://ftp.fil.ion.ucl.ac.uk/spm/data/MoAEpilot/ /MoAEpilot.tgz`.

[3] Frackowiak, R.S.J., Friston, K.J., Frith, C.D., Dolan, R.J., & Maziotta, J.C. (1997). Human Brain Function. *Academic Press*.

[4] Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., & Turner, R. (1998). Event-Related fMRI: Characterizing Differential Responses. *Neuroimage* , **7**, pp. 30 - 40.

[5] Gautama, T., & Van Hulle, M.M. (2000). Hierarchical Density-based Clustering In High-dimensional Spaces Using Topographic Maps. *IEEE Neural Network for Signal Processing Workshop 2000*, pp. 251 - 260.

[6] Gautama, T., & Van Hulle, M.M. (2001). Hierarchical Density-based Clustering of Shapes. *IEEE Neural Network for Signal Processing Workshop 2001, Massachusetts, in press*.

[7] Gold, S., Christian, B., Arndt, S., Zeien, G., Cizadlo, T., Johnson, D.L., Flaum, M. & Andreasen, N.C. (1998). *Human Brain Mapping*, **6**, pp. 73 - 84.

[8] Lars Kai Hansen, Finn Årup Nielsen, Peter Toft, Matthew G. Liptrot, Cyril Goutte, Stephen C. Strother, Nick Lange, Anders Gade, David A. Rottenberg, Olaf B. Paulson (1999), "lyngby" - a modeler's Matlab toolbox for spatio-temporal analysis of functional neuroimages. *NeuroImage*, **9** (6, part 2), pp. S241.

[9] Tononi, G., McIntosh, A.R., Russell, D.P. & Edelman, G.M. (1998). Functional clustering: identifying strongly interactive brain regions in neuroimaging data. *Neuroimage*, **7**, pp. 133 - 149.

[10] Van Hulle, M.M. (2000). Faithful Representations and Topographic Maps: From Distortion- to Information-based Self-Organization (Haykin, S. ed.). *John Wiley & Sons*.

DISCRIMINATIVE MINING OF GENE MICROARRAY DATA

Jianping Lu   Yue Wang   Zuyi Wang   Jianhua Xuan
San Yuan Kung   Zhiping Gu   Robert Clarke
The Catholic University of America, Washington, DC 20064, USA
Princeton University, Princeton, NJ 08544, USA
Celera Genomics, Inc., Rockville, MD 20850, USA
Georgetown University Medical Center, Washington, DC 20007, USA

Abstract. Spotted cDNA microarrays are emerging as a cost effective tool for the large scale analysis of gene expression. To reveal the patterns of genes expressed within a specific cell essentially responsible for its phenotype, this paper reports our progress in cluster discovery using a newly developed data mining method. The discussion entails: (1) statistical modeling of gene microarray data with a standard finite normal mixture distribution, (2) development of a joint supervised and unsupervised discriminative mining to discover sample clusters in a visual pyramid, and (3) evaluation of the data clusters produced by such scheme with phenotype-known microarray experiments.

INTRODUCTION

Spotted cDNA microarrays are emerging as a powerful and cost effective tool for the large scale analysis of gene expression. Using this technology, the relative expression levels in two or more mRNA populations derived from tissue samples can be assayed for thousands of genes simultaneously [1, 2]. Microarrays are potentially powerful tools for investigating the mechanism of drug action. Two recent studies have described the application of high density microarrays to examine the effects of drugs on gene expression in yeast as model system [1]. A similar method applied to human breast cancer cells and tissues would have direct utility in the identification and validation of novel therapeutics. It is widely accepted that the pattern of genes expressed within a specific cell is essentially responsible for its phenotype. The most widely publicized use of gene microarrays has been in cancer research [2].

For molecular analysis of cancer, the profile of microarray expression is the molecular signature of interest. The representation of each sample is described as a point in a d-dimensional gene expression space in which each axis represents the expression level of one gene. The presence of well-separated

0-7803-7196-8/01/$10.00 (C) 2001 IEEE

sample groups implies that the representations of samples within the same group are close to each other in this gene expression space but distant from those of other samples. Thus, the representations of phenotype-related samples form clusters. The research plan can be divided into three major steps: cluster discovery, gene selection, and phenotype prediction. Cluster discovery refers to detecting previously unrecognized tumor subtypes [2]. Gene selection refers to the identification of most relevant gene subset involving the biological process that generates the patterns. Phenotype prediction refers to the assignment of particular tumor sample to the known tumor classes which could indicate current states for future outcomes [2]. The main challenge is that the microarray data are high-dimensional, multimodal, and lacking in prior knowledge.

In this paper, we will report our progress in cluster discovery using a newly developed discriminative data mining method [3]. The presentation will entail three major issues: (1) statistical modeling of gene microarray data with a standard finite normal mixture (SFNM) distribution; (2) development of a joint supervised and unsupervised data mining scheme to "discover" sample clusters in a discriminative visual pyramid; and (3) evaluation of the data clusters produced by such scheme with phenotype-known microarray experiments. There are several major differences between our work and the previous most related research [4, 5, 6, 7, 8]. First, we developed a hierarchical visualization paradigm, involving mixture statistical submodels and visualization subspaces. The resulting data mining is capable of capturing all of the interesting aspects of the dataset, since the high complexity of data structure living in a high dimensional space cannot be adequately explored by a single-level visualization [4]. Secondly, we proposed a principle discriminative component analysis (PDCA) to probabilistically project the softly partitioned dataset onto multiple visual subspaces. It allows an effective separation of local clusters in dimensional reduced visual subspaces, rather than maximum likelihood which may well represent the original dataset but not necessarily good for cluster discovery [5, 8]. Furthermore, we implemented a probabilistic adaptive principal components extraction (PAPEX) algorithm to estimate the top two principal axes and an incremental expectation-maximization (IEM) procedure to estimation SFNM distribution. The computation is efficient when confronted with high dimensional datasets [10]. Finally, we imposed a model selection procedure to determine the number of subclusters within each cluster using the minimum description length criterion. This allows algorithm to automatically determine whether a further split of a subspace should continue in completing the whole hierarchy.

THEORY AND METHOD

Assume the sample points $\{\mathbf{t}_i\}$ in gene expression space form K_0 clusters $\{(\boldsymbol{\mu}_{\mathbf{t}1},\mathbf{C}_{\mathbf{t}1}),...,(\boldsymbol{\mu}_{\mathbf{t}k},\mathbf{C}_{\mathbf{t}k}),...,(\boldsymbol{\mu}_{\mathbf{t}K_0},\mathbf{C}_{\mathbf{t}K_0})\}$, where $\boldsymbol{\mu}_{\mathbf{t}k}$ and $\mathbf{C}_{\mathbf{t}k}$ are the mean vector and covariance matrix of cluster k respectively. Recently there has been considerable success in using the SFNM to model the distribution of a multimodal dataset [9], such that data distribution takes a sum of the following general form

$$p(\mathbf{t}) = \sum_{k=1}^{K_0} \pi_k g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k}) \qquad (1)$$

where π_k is the corresponding mixing proportion, with $0 \leq \pi_k \leq 1$ and $\sum \pi_k = 1$, and g is the Gaussian kernel. The problem of SFNM modeling applied to gene microarray data addresses the combined detection of structural parameter K_0 (e.g., cluster discovery) and estimation of regional parameters $(\pi_k, \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$, based on the observations $\mathbf{t}$. One natural criterion used for the modeling is the maximum likelihood (ML) estimation using the expectation-maximization (EM) algorithm [9].

Since the dimensionality of gene microarray data is very high ($500 \sim 8000$), three challenging problems are associated with the current approaches [3]. First, the number of the local clusters in a particular dataset is generally unknown, model selection is a prerequisite. Second, the computational complexity of the EM algorithm running in $\mathbf{t}$-space is accordingly high. Third, the initialization of the EM algorithm is often blindly or heuristically chosen, which may lead to both local optima and slow convergence. A natural way, with greater practical applicability, to tackle these problems is to introduce user interaction with the data mining system. For example, by examining plots of principal component space, researchers often develop a deeper understanding of the driving forces that generated the original data, and effortlessly grasp the general characteristics of the data and propose an initial solution.

One of the difficulties inherent in visual mining is the problem of visualizing multi-dimensionality [4]. When there are more than three variables, it stretches the imagination to visualize their relationships. By fully taking advantage of the multimodal nature of gene microarray data, our approach for hierarchical cluster discovery includes two complementary components: (1) dimensionality reduction by probabilistic principal component projection and (2) cluster decomposition by adaptive soft data partitioning.

Principal component analysis (PCA) is an effective unsupervised method for achieving dimensionality reduction [6, 10, 11]. For a set of observed d-dimensional data vectors $\{\mathbf{t}_i\}$, $i \in \{1,...,N\}$, the q principal axes $\mathbf{w}_m$, $m \in \{1,...,q(\leq d)\}$, are those orthogonal axes onto which the retained variance under projection is maximal. It can be shown that the principal axes $\mathbf{w}_m$ are given by the q dominant eigenvectors (i.e., maximal eigenvalues) of the sample covariance matrix $\mathbf{C}_{\mathbf{t}}=\frac{1}{N}\sum_{i=1}^{N}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})^{\mathrm{T}}$ such that

$$\mathbf{C}_{\mathbf{t}}\mathbf{w}_m = \lambda_m \mathbf{w}_m \qquad (2)$$

where $\boldsymbol{\mu}_{\mathbf{t}}$ is the sample mean and λ_m is the eigenvalue. The vector $\mathbf{x}_i = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}})$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_q)$, is thus a q dimensional new representation of the observed vector $\mathbf{t}_i$.

Two related issues contribute to the limitations of conventional PCA: its global linearity without incorporating data structure; and its optimality based on reconstruction error rather than pattern separability. Thus, genes with large expression variances but less discriminant abilities may play dominant roles in determining the projections. A more effective way when confronted with a multimodal dataset, however, is to emphasize the inter-cluster separation by replacing the total covariance matrix with the Fisher's scatter matrix [5, 11], i.e., to find the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{w}_m = \lambda_m \mathbf{w}_m \tag{3}$$

where the *within-cluster scatter matrix* ($\mathbf{S}_w$) is the *joint* scatter of data point $\mathbf{t}_i$ around the conditional mean vector $\boldsymbol{\mu}_{\mathbf{t}k}$ of different clusters, i.e., $\mathbf{S}_w = \sum_{k=1}^{K_0} \pi_k \mathbf{C}_{\mathbf{t}k}$ with cluster conditioned covariance matrix $\mathbf{C}_{\mathbf{t}k} = \sum_{i=1}^{N} z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})^{\mathrm{T}} / \sum_{i=1}^{N} z_{ik}$ and $z_{ik} = \pi_k g(\mathbf{t}_i|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})/p(\mathbf{t}_i)$, and the *between-cluster scatter matrix* ($\mathbf{S}_b$) is the scatter of the cluster conditional mean vector $\boldsymbol{\mu}_{\mathbf{t}k}$ around the overall data center $\boldsymbol{\mu}_{\mathbf{t}}$, i.e., $\mathbf{S}_b = \sum_{k=1}^{K_0} \pi_k(\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}})(\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}})^{\mathrm{T}}$, such that the separability of patterns is maximized, that is

$$\mathbf{W} = \arg\max_{\mathbf{W}_0}\{Trace(\mathbf{W}_0^{\mathrm{T}}\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{W}_0)\}. \tag{4}$$

This is termed as the principal discriminative component analysis (PDCA).

Now consider a two-dimensional projection space $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ together with a linear transformation, that maps the data space to the projection space by $\mathbf{x} = \mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu}_{\mathbf{t}})$ where $\mathbf{W}$ is a $d \times 2$ matrix. For a normal distribution $g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$ over the data space, a similar reduced dimension probability distribution $g(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$ of the new variable $\mathbf{x}$ in the projection space is simply defined by the *Radon* transform of $g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$, i.e., $g(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}k}, \mathbf{C}_{\mathbf{x}k}) = \int g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})\delta(\mathbf{x} - \mathbf{W}^{\mathrm{T}}\mathbf{t} + \mathbf{W}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}})dt$ where $\delta(.)$ is the delta function that $\delta(0) = 1$ and $\delta(\neq 0) = 0$. According to the linear superposition property of *Radon* transform and the projection invariant property of normal distribution, we have

$$f(\mathbf{x}) = \sum_{k=1}^{K_0} \pi_k g(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}k}, \mathbf{C}_{\mathbf{x}k}). \tag{5}$$

as the counterpart of Eq. (1) in $\mathbf{x}$-space defined by projection matrix $\mathbf{W}$.

However, when dataset is projected onto a single lower dimensional subspace, its inherent multimodal nature may be partially or completely obscured according to Cover's theorem on the separability of patterns [10]. In other words, even though the cluster structure of a dataset may be evident from the higher dimensional plot, it is quite conceivable to have the finer cluster patterns concealed after a single linear projection, leading to an unidentifiable correspondence between Eq. (1) and Eq. (5) [3]. A novel approach is to model high-dimensional multimodal dataset with a hierarchical mixture model and accordingly with a collection of probabilistic principal discriminative subspaces [3, 6, 7, 8], namely the exploratory cluster discovery.

Assume a top-level model consisting of a single *Radon* transform $\mathbf{W}$ and a mixture of $K_1(< K_0)$ normal distributions $p(\mathbf{t}) = \sum_{k=1}^{K_1} \pi_k g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$ which is identifiable in $\mathbf{x}$-space, i.e., $f(\mathbf{x}) = \sum_{k=1}^{K_1} \pi_k g(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}k}, \mathbf{C}_{\mathbf{x}k})$, we can form a two-level hierarchy by associating a group of SFNM submodels with each model k at top-level

$$p(\mathbf{t}) = \sum_{k=1}^{K_1} \pi_k \sum_{j=1}^{K_{2,k}} \pi_{j|k} g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)}) \tag{6}$$

where $\pi_{j|k}$ again corresponds to a set of mixing proportions, one for each k, with $0 \leq \pi_{j|k} \leq 1$ and $\sum_j \pi_{j|k} = 1$, and $\sum_{k=1}^{K_1} K_{2,k} = K_0$. To reveal the hidden cluster pattern within each model k at top-level, i.e., $g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k}) = \sum_{j=1}^{K_{2,k}} \pi_{j|k} g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})$, an associated probabilistic principal discriminative subspace is constructed which focuses on the separability of patterns within the data portion defined by model k, where the opaque degree of a data point in the subspace plot is proportional to its posterior probability belonging to this model, i.e., z_{ik} determined at top-level.

The further cluster discovery is a two-stage procedure: a soft partitioning of each model k into $K_{2,k}$ subclusters followed by a construction of corresponding subspace. Instead of any given data point being assigned exclusively to one subspace, the responsibility for its generation is shared among all of the subspaces. The subspaces for the submodels at second-level are generated by the probabilistic PDCA such that

$$\mathbf{S}_{k,w}^{-1}\mathbf{S}_{k,b}\mathbf{w}_{k,m} = \lambda_{k,m}\mathbf{w}_{k,m} \tag{7}$$

where $\mathbf{S}_{k,w} = \sum_{j=1}^{K_{2,k}} \pi_{j|k}\mathbf{C}_{\mathbf{t}(k,j)}$ with subcluster conditioned covariance matrix $\mathbf{C}_{\mathbf{t}(k,j)} = \sum_{i=1}^{N} z_{i(k,j)}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})^{\mathrm{T}} / \sum_{i=1}^{N} z_{i(k,j)}$ and $z_{i(k,j)} = z_{ik}\pi_{j|k} g(\mathbf{t}_i|\boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})/g(\mathbf{t}_i|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$, and $\mathbf{S}_{k,b} = \sum_{j=1}^{K_{2,k}} \pi_{j|k}(\boldsymbol{\mu}_{\mathbf{t}(k,j)} - \boldsymbol{\mu}_{\mathbf{t}k})(\boldsymbol{\mu}_{\mathbf{t}(k,j)} - \boldsymbol{\mu}_{\mathbf{t}k})^{\mathrm{T}}$. The probability distribution of model k in $\mathbf{x}$-space at second-level is now defined by the model k focused *Radon* transform of $g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$, i.e., $g(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}k}, \mathbf{C}_{\mathbf{x}k}) = \int g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})\delta(\mathbf{x} - \mathbf{W}_k^{\mathrm{T}}\mathbf{t} + \mathbf{W}_k^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{t}k})dt$. It should be noted that each component in Eq. (6) now corresponds to an independent subspace model with projection matrix $\mathbf{W}_k$. To interpret the corresponding set of visualization subspaces, all data points $\mathbf{x}_{ik} = \mathbf{W}_k^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})$ will appear in every plot of the total K_1 subspaces at the second-level, with their opaque degree-equal or proportional to z_{ik}.

The hierarchical version of the SFNM model can be further extended to include more levels based on the same principle as above. The deeper the tree is, the more submodels are used and the finer are these submodels. The formation of the hierarchy is guided by model selection over $\mathbf{x}$-subspaces. Each level of the hierarchy corresponds to a generic subspace, with lower levels giving more focused visual interpretations. Model selection refers to the detection of the structural parameter K. In addition to user's visual inspection, we propose to use an information theoretic criterion, i.e., the *minimum description length* (MDL) [12], to guide model selection. The major thrust of this approach has been the formulation of a model fitting procedure in which an optimal model is selected from the several competing candidates such that the selected model best fits the observed data. Thus, the value of K is selected by minimizing

$$\mathrm{MDL}(K_a) = -\log(\mathcal{L}_{ML}) + 0.5K_a \log N \tag{8}$$

where K_a is the number of free adjustable parameters, and $\mathcal{L}_{ML}$ is the joint maximum likelihood, of the model respectively.

NEURAL COMPUTATION

We now present the description of our algorithm. Data mining is not process that can be orchestrated a priori, and knowledge discovery must follow for insight and spontaneous inspiration. Our algorithm progressively proceeds by fitting a series of submodels to the clusters of the dataset *interactively* and *incrementally.*

Our algorithm begins by determining $\mathbf{W}$ for the top-level projection (a single two-dimensional $\mathbf{x}$-space). The initial estimate of $\mathbf{W}$ is obtained from our previously developed APEX neural computation (e.g., Eq. (2)) [10], and further modified by PDCA-APEX algorithm (e.g., Eq. (3) with or without $\mathbf{S}_w^{-1}$) [3] where the prerequisite is to estimate the SFNM model at top-level (e.g., Eq. (1)). To remedy the problem of high dimensionality with gene microarray data, neural computation of $(\mathbf{W}, \pi_k, \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})$ is efficient in that only the top two eigenvectors of the covariance or scatter matrix are calculated, and model parameter values are estimated firstly in $\mathbf{x}$-space and further fine tuned in $\mathbf{t}$-space incrementally. For example, the IEM procedure provides "soft" splits of the data points, hence allowing the data to contribute simultaneously to multiple clusters which results in

E-Step

$$z_{(i+1)k} = \frac{\pi_k^{(i)} g(\mathbf{x}_{i+1} | \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}, \mathbf{C}_{\mathbf{x}k}^{(i)})}{f(\mathbf{x}_{i+1} | \pi_k^{(i)}, \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}, \mathbf{C}_{\mathbf{x}k}^{(i)})}, \tag{9}$$

M-Step

$$\boldsymbol{\mu}_{\mathbf{x}k}^{(i+1)} = \boldsymbol{\mu}_{\mathbf{x}k}^{(i)} + a(i)(\mathbf{x}_{i+1} - \boldsymbol{\mu}_{\mathbf{x}k}^{(i)}) z_{(i+1)k}, \tag{10}$$

$$\mathbf{C}_{\mathbf{x}k}^{(i+1)} = \mathbf{C}_{\mathbf{x}k}^{(i)} + b(i)[(\mathbf{x}_{i+1} - \boldsymbol{\mu}_{\mathbf{x}k}^{(i)})(\mathbf{x}_{i+1} - \boldsymbol{\mu}_{\mathbf{x}k}^{(i)})^{\mathrm{T}} - \mathbf{C}_{\mathbf{x}k}^{(i)}] z_{(i+1)k}, \tag{11}$$

$$\pi_k^{(i+1)} = \frac{i}{i+1}\pi_k^{(i)} + \frac{1}{i+1} z_{(i+1)k} \tag{12}$$

for $k = 1, ..., K_1$, where $a(i)$ and $b(i)$ are introduced as the learning rates, two sequences converging to zero, ensuring unbiased estimates after convergence. The user will pin-point the initial cluster centers $\boldsymbol{\mu}_{\mathbf{x}k}^{(0)}$ and assign $\pi_k^{(0)} = 1/K_1$ and $\mathbf{C}_{\mathbf{x}k}^{(0)} = \mathbf{W}^{\mathrm{T}}\mathbf{C}_{\mathbf{t}}\mathbf{W}$. The optimum value of K_1 is determined based on MDL (e.g., Eq. (8)) where $K_a = 6K_1 - 1$.

Determination of the subspaces $\mathbf{W}_k$ and submodels $(\pi_{j|k}, \boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})$ at the second-level can again be viewed as a two-step estimation problem, in which further split of the models is determined within each of the clusters identified at the top-level such that its internal structures can be further explored over cluster-focused $\mathbf{x}$-space. The initial estimate of $\mathbf{W}_k$ can be obtained using a probabilistic APEX (PAPEX) algorithm as we now sketch:

1. Initialize the feedforward weight vector $\mathbf{w}_{k,m}$ for $m = 1, 2$, and the feedback weight vector a_k to small random values at time $i = 1$. Assign a small positive value to the learning rate parameter η.

2. Set $m = 1$, and for $i = 1, 2, ...,$ compute

$$y_{k,1}(i) = \mathbf{w}_{k,1}^{\mathrm{T}}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}) \tag{13}$$

$$\mathbf{w}_{k,1}(i+1) = \mathbf{w}_{k,1}(i) + \eta[y_{k,1}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}) - y_{k,1}^2(i)\mathbf{w}_{k,1}(i)] \tag{14}$$

For large i we have $\mathbf{w}_{k,1}(i) \longrightarrow \mathbf{w}_{k,1}$, where $\mathbf{w}_{k,1}$ is the eigenvector associated with the largest eigenvalue of the covariance matrix $\mathbf{C}_{\mathbf{t}k}$.

3. Set $m = 2$, and for $i = 1, 2, ...,$ compute

$$y_{k,2}(i) = \mathbf{w}_{k,2}^{\mathrm{T}}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}) + a_k(i) y_{k,2}(i) \tag{15}$$

$$\mathbf{w}_{k,2}(i+1) = \mathbf{w}_{k,2}(i) + \eta[y_{k,2}(i) z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k}) - y_{k,2}^2(i)\mathbf{w}_{k,2}(i)] \tag{16}$$

$$a_k(i+1) = a_k(i) - \eta[y_{k,2}(i) y_{k,2}(i) + y_{k,2}^2(i) a_k(i)] \tag{17}$$

For large i we have $\mathbf{w}_{k,2}(i) \longrightarrow \mathbf{w}_{k,2}$, where $\mathbf{w}_{k,2}$ is the eigenvector associated with the second largest eigenvalue of the covariance matrix $\mathbf{C}_{\mathbf{t}k}$.

The corresponding $\sum_{j=1}^{K_{2,k}} \pi_{j|k} g(\mathbf{t}|\boldsymbol{\mu}_{\mathbf{t}(k,j)}, \mathbf{C}_{\mathbf{t}(k,j)})$ can again be estimated using IEM algorithm to allow a SFNM distribution with $K_{2,k}$ submodels to be fitted to cluster k, where the user will pin-point the initial subcluster centers $\boldsymbol{\mu}_{\mathbf{x}(k,j)}^{(0)}$ and assign $\pi_{j|k}^{(0)} = 1/K_{2,k}$ and $\mathbf{C}_{\mathbf{x}(k,j)}^{(0)} = \mathbf{W}_k^{\mathrm{T}}\mathbf{C}_{\mathbf{t}k}\mathbf{W}_k$ to initialize $\sum_{j=1}^{K_{2,k}} \pi_{j|k} g(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}(k,j)}, \mathbf{C}_{\mathbf{x}(k,j)})$ with a model selection procedure. By replacing $z_{ik}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}k})$ in PAPEX formulation with $\pi_{j|k}(\boldsymbol{\mu}_{\mathbf{t}(k,j)} - \boldsymbol{\mu}_{\mathbf{t}k})$, $\mathbf{W}_k$ is updated by a PDCA-PAPEX procedure to generate a separability-based and cluster-focused subspace for model k at the second-level.

The construction of the entire tree structure hierarchy is completed when no further data split is recommended in all of the parent subspaces, followed by the generation of the bottom-level subspaces (for example, the third-level). The value of $\mathbf{W}_{(k,j)}$ is obtained using the PAPEX algorithm with $z_{i(k,j)}$ instead of z_{ik}, and all data points $\mathbf{x}_{i(k,j)} = \mathbf{W}_{k,j}^{\mathrm{T}}(\mathbf{t}_i - \boldsymbol{\mu}_{\mathbf{t}(k,j)})$ will appear in every plot of the total K_0 subspaces at the bottom-level, with their opaque degree equal or proportional to $z_{i(k,j)}$.

We first illustrate the application of our *VISual Data Analysis* (VISDA) algorithm to a synthetic Toy dataset provided in [8] "consisting of 450 data points generated from a mixture of three Gaussians in three-dimensional space. Each Gaussian is relatively flat (has small variance) in one dimension. Two of these pancake-like clusters are closely spaced, while the third is well separated from the first two." After the initial top-level projection generated by a global PCA, shown in Fig. 1 (Left), the result of model selection suggested the presence of two distinct clusters. User interaction and PDCA are then applied to construct subspaces at the second-level. The information theoretic criterion suggested a further split of cluster 2 (blue&yellow) but not of cluster 1 (red). The final cluster visualization is completed similarly through three corresponding cluster-focused subspaces at the bottom-level.

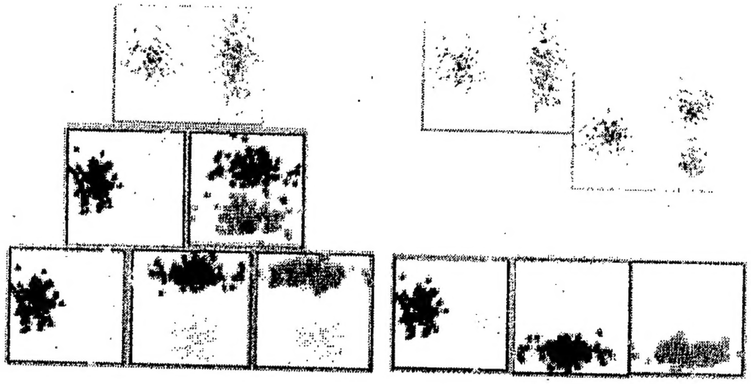


Figure 1: Result of incremental cluster discovery using VISDA applied to Toy Data (Left: three-level decomposition. Right: two-level decomposition.).

With this three-level cluster discovery, the basic principle and capable nature of the approach are evident as the interim subspaces (second-level) attempt to reveal hidden clusters which have been insufficiently modeled by their immediate ancestor (top level). The original data clusters have been individually colored, and the procedure successfully discovered all three data clusters. An alternative discovery pathway is shown in Fig. 2 (Right) where PDCA iteration driven by the user discovered all three data clusters at the top-level and reached the same result as through the first pathway.

As an example of more realistic problems, we considered Iris dataset consisting of 150 four-dimensional points in three classes: Setosa, Versicolor, and Virginica. The same cluster discovery pathways used for analyzing Toy dataset were applied to this dataset. The result indicated a successful application of the new approach to real-world problems, as shown in Fig. 2. All three data clusters have been discovered and highlighted. The PDCA mapping makes use of the perceivable cluster structure at each level, this being the key reason why this element leads to a discriminative cluster discovery.

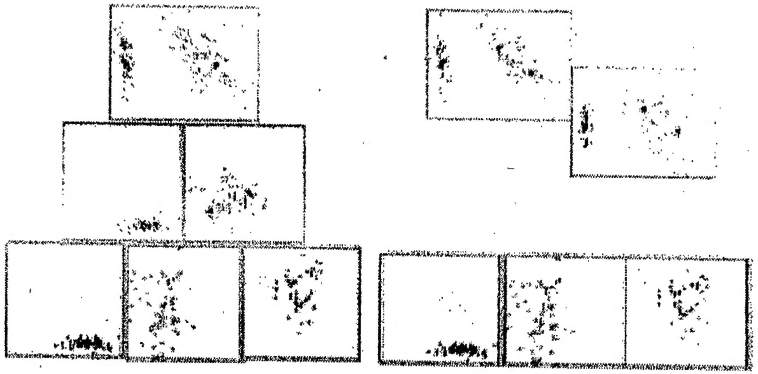


Figure 2: Result of incremental cluster discovery using VISDA applied to Iris Data (Left: three-level decomposition. Right: two-level decomposition.).

The final example reports the effectiveness of VISDA in cluster discovery with real gene microarray data. The dataset consists of 72 samples with 47 of acute lymphoblastic leukemia (ALL) and 25 of acute myeloid leukemia (AML), provided in [2] where distinguishing ALL from AML is critical for successful treatment. The original 6817 genes evaluated using Affymetrix cDNA microarrays were ranked by their differentiation ability between ALL and AML and the top 1100 genes were retained for further analysis.

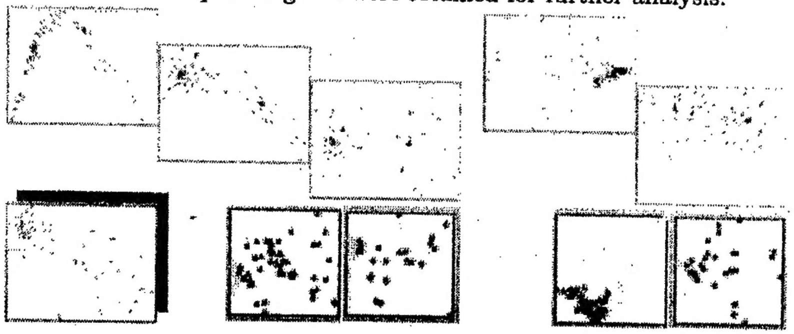


Figure 3: Summary of exploratory cluster discovery using VISDA applied to ALL-AML Data (Left: subspaces of top 22 genes. Right: subspaces of other genes.).

To explore whether the ALL-AML distinction could have been discovered blindly on the basis of informative gene expression, we applied VISDA to this dataset using the top 22 genes. Fig. 3 (Left) shows the result of iterative PDCA exploration with a two-level pyramid. We first evaluated the clusters by comparing them to the known ALL-AML classes. It can be seen that after the first two iterations, blind cluster discovery (3rd subplot of the top-level) paralleled the known classes (subplot with shadow) closely. An independent trial using other genes reached a similar result but less discriminative, shown in Fig. 3 (Right). The data clusters associated with the ALL-AML were further characterized closely through the bottom-level subspaces. However, VISDA did not reveal the finer subclusters corresponding to T-cell and B-cell ALL as reported in [2]. With larger sample collections, it would be possible to define or validate new cancer subtypes.

While the optimality of new techniques is often highly data-dependent, we would expect VISDA to be a very effective tool in gene microarray data analysis. We are currently investigating further applications to the molecular classification of breast cancer.

REFERENCES

[1] D. J. Duggan, M. L. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nature*, vol. 21, pp. 10-14, Jan. 1999.

[2] T. R. Golub, D. K. Slonim, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, Oct. 1999.

[3] Y. Wang, L. Luo, M. T. Freedman, and S-Y Kung, "Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization," *IEEE Trans. Neural Nets*, vol. 11, no. 3, pp. 625-636, May 2000.

[4] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, Dec. 2000.

[5] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 6, pp. 623-627, June 2000.

[6] G. E. Hinton, P. Dayan, and M. Revow, "Modeling the manifolds of images of handwritten digits," *IEEE Trans. Neural Net.*, vol. 8, no. 1, pp. 65-74, Jan. 1997.

[7] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493-1516, 1997.

[8] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, pp. 443-482, 1999.

[9] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical analysis of finite mixture distributions*. New York: John Wiley, 1985.

[10] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice-Hall, Inc., Upper Saddle River, Ney Jersey, 1999.

[11] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[12] J. Rissanen, "Modeling by shortest data description," *Automat.*, vol. 14, pp. 465-471, 1978.

A LOCAL WEIGHTING METHOD TO THE INTEGRATION OF NEURAL NETWORK AND CASE BASED REASONING

Jae Heon Park[a], Chung-Kwan Shin[b], Kwang Hyuk Im[c], and Sang Chan Park[d]
[a,c,d] Department of Industrial Engineering,
Korea Advanced Institute of Science and Technology, Taejon, Korea
[b] Electronics and Telecomminucations Research Institute(ETRI), Taejon, Korea
Phone, Fax: +82 42 869 2960
E-mail: [a]dewy@major.kaist.ac.kr, [b]ckshin@etri.re.kr, [c]gunni@major.kaist.ac.kr, [d]sangpark@kaist.ac.kr

Abstract. Our aim is to build an integrated learning framework of neural network and case based reasoning. The main idea is that feature weights for case based reasoning can be evaluated using neural networks. In our previous method, we derived the feature weight set from the trained neural network and the training data so that the feature weight is constant for all queries. In this paper, we propose a local feature weighting method using a neural network. The neural network guides the case based reasoning by providing case-specific weights to the learning process. We developed a learning process to get the local weights using the neural network and showed the performance of our learning system using the sinusoidal dataset.

INTRODUCTION

Our aim is to build the method of integrating NN (Neural Network) and CBR (Case Based Reasoning) in a local weighting approach. In our method, a neural network is trained to guide and coordinate the reasoning process in CBR. When a new query is given, the trained neural network adjusts distances between the query and cases in the case base considering the position of the query in the case space. The structure of the neural network is designed to learn the characteristics of input features. The trained neural network adjusts feature weights by calculating the distance between the query and cases in the case base.

In the second section, we introduce our previous global weighting methods using neural network. In the third section, the proposed system and the training algorithm are introduced. Then we show the experimental results of our system using the sinusoidal dataset.

GLOBAL FEATURE WEIGHTING METHOD BY NEURAL NETWORK

Shin et al [1] proposed a hybrid system of neural network and memory-based reasoning. This system approaches the comprehensible knowledge problem of

0-7803-7196-8/01/$10.00 (C) 2001 IEEE